



MADURAI KAMARAJ UNIVERSITY

(University with Potential for Excellence)

DISTANCE EDUCATION

Recognised by DEC

www.mkudde.org

B.Sc.,

MATHEMATICS

Second Year

Paper - IV

STATISTICS

Volume 1

Units : 1 - 5

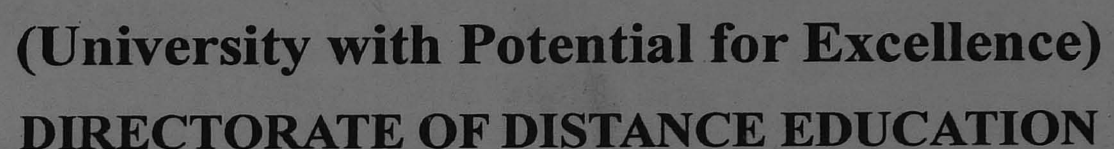




S 85



MADURAI KAMARAJ UNIVERSITY



(University with Potential for Excellence)

DIRECTORATE OF DISTANCE EDUCATION



MATHEMATICS

B.Sc (Mathematics)
Second Year

PAPER - IV

STATISTICS

Volume - I : 1 to 5 Units

S 85

B.Sc.,(Mathematics)

Second Year

WELCOME

Dear students,

We welcome you as a student of the first year B.Sc., degree course.

This book deals with the subject **STATISITICS**. The learning material for this book will be supplemented by contact lectures.

In this book, first part deals central tendency, measure of dispersion and correlations. Later part deals sampling theory which is important in our day to applications.

Learning through the Distance Education mode, as you are all aware, involves self-learning and self-assessment and in this regard, you are expected to part in disciplined and dedicated effort.

As our part, we assure of our guidance and support.

With best regards.

STATISTICS

Unit I :

Measures of averages – Measures of dispersion – Skewness based on moments.

Unit 2 :

Correlation and Regression – Rank correlation coefficient.

Unit 3:

Index numbers and Time series.

Unit 4 :

Curve fitting (All types of curves)

Unit -5 :

Theory of Attributes.

Unit 6 :

Theory of probability – sample space- probability function – Laws of Addition – Boole's inequality – Law of Multiplication – Problems – Baye's theorem – Problems.

Unit 7 :

Random variables – Distribution function – Discrete and Continuous random variables – Probability density function – Mathematical Expectations (One dimension only).

Unit 8 :

Moment generating function – cumulants – Theoretical distributions – binomial, Poisson, Normal.

Unit 9 :

Tests of significance of Large Samples.

Unit 10 :

Tests of significance of small samples – t, F, χ^2.

Text Book :

Statistics by Dr. S.Arumugam – Sci-Tech Publications, 2006.

B. Sc MATHEMATICS – FIRST YEAR

PAPER – IV

STATISTIC

SCHEME OF LESSONS

UNIT – X

Tables

B.Sc.,(Mathematics)

Second Year

WELCOME

Dear students,

We welcome you as a student of the first year B.Sc., degree course.

This paper deal with the subject **STATISITICS**. The learning material for this paper will be supplemented by contact lectures.

In this book, first part deals central tendency, measure of dispersion and correlations. Later part deals sampling theory which is important in our day to applications.

Learning through the Distance Education mode, as you are all aware, involves self-learning and self-assessment and in this regard, you are expected to part in disciplined and dedicated effort.

As our part, we assure of our guidance and support.

With best regards.

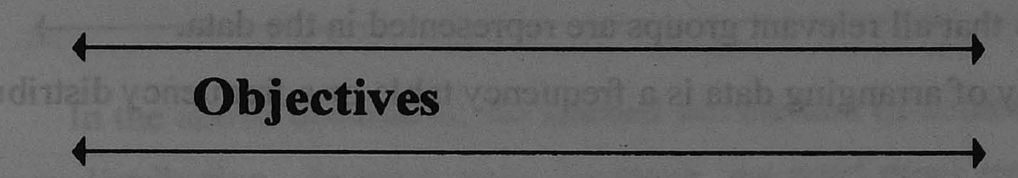
Lesson writer :
N.H.SARAVANAN,
M.Sc., M.Phil., PGDCA., PGDOR.,
PG.Dip. in Econometics.,
Associate Professor,
Dept. of MATHEMATICS,
SOURASHTRA COLLEGE,
MADURAI – 625 004.

Dr. V.KALAIMANI,
Professor and Head,
Dept. of MATHEMATICS,
Directorate of Distance Education,
Madurai Kamaraj University,
Madurai – 625 021.

Space for Hint

Unit I

Measures of averages and dispersions



Objectives

In this unit, we are going to discuss mean, median, mode, geometric mean, harmonic mean, range, mean deviation, quartile deviation, standard deviation, and skewness.

After the completion of this unit one may able to find

- Mean
- Median
- Mode
- Geometric mean
- Harmonic mean
- Range
- Mean Deviation
- Quartile Deviation
- Standard Deviation
- Skewness

Introduction

Statistics deals with collection, classification, tabulation, analysis, and interpretation of numerical data.

The word *statistic* comes from the Italian word *statista* (meaning *statesman*). It was first used by *Gottfried Achenwall* (1719 – 1772), a professor at Marlborough and Götingen. Dr. E.A.W.Zimmerman introduced the word *statistics* into England. Sir John Sinclair popularized its use in his work Statistical Account of Scotland 1791 – 1799.

Data are collections of any no of related observations. We can collect mathematics marks of students who were joined in higher secondary class.

Space for Hint

For data to be useful, our observations must be organized so that we can pick out patterns and come to logical conclusions. Statisticians select their observations so that all relevant groups are represented in the data.

One way of arranging data is a frequency table or a frequency distribution.

For example, consider the following raw data which gives the marks of 100 students.

18	86	14	31	67	45	36	71	30	41
44	66	31	81	90	39	37	61	56	50
41	70	61	58	40	26	07	16	13	08
66	20	81	91	71	77	66	38	41	47
50	61	64	38	61	39	48	78	70	90
66	38	17	66	26	98	16	10	09	61
21	44	33	21	74	68	72	66	03	18
61	53	55	41	70	39	30	06	02	81
91	03	13	18	20	30	31	46	48	11
50	33	41	48	67	69	38	40	07	10

In the above data, as it is, we cannot find any information. A better way is to arrange the data in the ascending or descending order of magnitude. Data expressed in ascending or descending order is called *data array*. Further, after arranging into ascending or descending order it is difficult to draw valid or meaningful conclusions. Thus we can arrange the data in a frequency distribution, which is shown below.

Class Interval	No. of students
0 - 10	8
10 - 20	12
20 - 30	6
30 - 40	17
40 - 50	15
50 - 60	7
60 - 70	17
70 - 80	9
80 - 90	4
90 - 100	5

The above table is called frequency distribution.

Space for Hint

1. 1 Central Tendencies

In the above discussion, we learned the method of construction of frequency distribution. In most cases, however, we need more exact measures. We can use simple numbers called *summary statistics* to describe characteristics of a data set.

Two of these characteristics are popularly important to decision makers, namely *central tendency* and *dispersion*.

Central tendency is the middle point of a distribution. Measures of central tendency are also called *measures of location*. The following are the five measures of central tendency which are in common use.

1. Arithmetic mean
2. Median
3. Mode
4. Geometric Mean
5. Harmonic Mean

1.1. 1 Arithmetic Mean

Definition : The arithmetic mean of n observations $x_1, x_2, x_3, \cdots, x_n$ is defined

by $\bar{x} = \dfrac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$

(i.e) $\bar{x} = \dfrac{\Sigma x}{n}$

Definition : Let $x_1, x_2, x_3, \cdots, x_n$ having frequencies $f_1, f_2, f_3, \cdots, f_n$ then the arithmetic mean of the distribution is $\bar{x} = \dfrac{f_1x_1 + f_2x_2 + f_3x_3 + \cdots + f_nx_n}{f_1 + f_2 + f_3 + \cdots + f_n}$

(i.e) $\bar{x} = \dfrac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$

Space for Hint

(i.e) $\overline{x} = \dfrac{\sum fx}{N}$ where $N = \sum f_i = f_1 + f_2 + f_3 + \cdots + f_n$

Definition : If $x_1, x_2, x_3, \cdots, x_n$ having weights $w_1, w_2, w_3, \cdots, w_n$, then the *weighted mean* or *weighted average* is given by

$$\overline{x} = \frac{w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_nx_n}{w_1 + w_2 + w_3 + \cdots + w_n}$$

$$\text{(i.e)} \quad \overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}.$$

Exapmle 1. 1

Find the average of 20, 17, 20, 18, 19, 20, 17, 21, 19, 19.

Solution : Average $= \dfrac{20+17+20+18+19+20+17+21+19+19}{10}$

$= \dfrac{190}{10}$

$= 19$

Exapmle 1. 2

Calculate the mean from the following data :

x :	1	2	3	4	5	6
f :	17	21	36	27	7	3

Solution :

We know that mean $= \overline{x} = \dfrac{\sum fx}{N}$

x	f	fx
1	17	17
2	21	42
3	35	105
4	27	108
5	7	35
6	3	18
Total	110	325

$$\therefore\ \bar{x} = \frac{\sum fx}{N}$$

$$= \frac{325}{110}$$

$$= 2.9545$$

Exapmle 1. 3

The following table gives the male population (in Lakhs) of a city in a certain year. Find the mean age of the males.

Age group	No. of males
0 - 5	14
5 - 10	13
10 - 15	13
15 - 20	13
20 - 30	33
30 - 40	29
40 - 50	17
50 - 60	07
60 - 80	04

Solution :

We know that $\bar{x} = \frac{\sum fx}{N}$

Age group	No. of males (f)	*mid-value* (x)	fx
0 - 5	14	2.5	35.0
5 - 10	13	7.5	97.5
10 - 15	13	12.5	162.5
15 - 20	13	17.5	227.5
20 - 30	33	25.0	825.0
30 - 40	29	35.0	1015.0
40 - 50	17	45.0	765.0
50 - 60	07	55.0	385.0
60 - 80	04	70.0	280.0
Total	143		3792.5

Space for Hint

Space for Hint

$$\therefore \ \bar{x} = \frac{\sum fx}{N}$$

$$= \frac{3792.5}{143}$$

$$= 26.521$$

Note : The formula of $\bar{x} = \frac{\sum fx}{N}$ can also be modified as $\bar{x} = A + \frac{\sum f_i d_i}{N}$

where $d_i = x_i - A$ and A is called *assumed mean.*

Now we shall solve the above example 1.3 using the above formula

Age group	No. of males (f)	*mid-value* (x)	$d = x - A$	fx
0 - 5	14	2.5	-22.5	-315.0
5 - 10	13	7.5	-17.5	-227.5
10 - 15	13	12.5	-12.5	-162.5
15 - 20	13	17.5	-7.5	-97.5
20 - 30	33	25.0	0.0	0.0
30 - 40	29	35.0	10.0	290.0
40 - 50	17	45.0	20.0	340.0
50 - 60	07	55.0	30.0	210.0
60 - 80	04	70.0	45.0	180.0
Total	143			217.5

$$\therefore \ \bar{x} = A + \frac{\sum f_i d_i}{N}$$

$$= 25 + \frac{217.5}{143}$$

$$= 26.521$$

Exapmle 1. 4 :

Given the following frequency distribution, calculate arithmetic mean.

Space for Hint

Daily wages in (Rs.)	No. of workers (f)
12.5 - 17.5	20
17.5 - 22.5	220
22.5 - 27.5	100
27.5 - 32.5	140
32.5 - 37.5	30
37.5 - 42.5	40
42.5 - 47.5	60
47.5 - 52.5	10
52.5 - 57.5	20

Solution :

We know that $\overline{x} = A + \frac{\sum f_i d_i}{N}$

Daily wages in (Rs.)	No. of workers (f)	*mid-value* (x)	$d = x - A$	fx
12.5 - 17.5	20	15.0	-20.0	-400.0
17.5 - 22.5	220	20.0	-15.0	-3300.0
22.5 - 27.5	100	25.0	-10.0	-1000.0
27.5 - 32.5	140	30.0	-5.0	-700.0
32.5 - 37.5	30	35.0	0.0	0.0
37.5 - 42.5	40	40.0	5.0	200.0
42.5 - 47.5	60	45.0	10.0	600.0
47.5 - 52.5	10	50.0	15.0	150.0
52.5 - 57.5	20	55.0	20.0	400.0
Total	640			-4050.0

We know that $\overline{x} = A + \frac{\sum f_i d_i}{N}$

Choose $A = 35$

$\therefore \quad \overline{x} = A + \frac{\sum f_i d_i}{N}$

Space for Hint

(i.e) $\bar{x} = 35 + \frac{(-4050)}{640}$

(i.e) $\bar{x} = 35 + \frac{(-4050)}{640}$

(i.e) $\bar{x} = 35 - 6.328$

(i.e) $\bar{x} = 28.672$

Note : the formula $\bar{x} = A + \frac{\sum f_i d_i}{N}$ can be modified as $\bar{x} = A + \frac{\sum f_i d_i}{N} \times c$

where $d_i = \frac{x_i - A}{c}$

We use the above formula when all the class intervals have the same length.

Daily wages in (Rs.)	No. of workers (f)	*mid-value* (x_i)	$d_i = \frac{x_i - 35}{5}$	$f_i x_i$
12.5 - 17.5	20	15.0	-4.0	-80.0
17.5 - 22.5	220	20.0	-3.0	-660.0
22.5 - 27.5	100	25.0	-2.0	-200.0
27.5 - 32.5	140	30.0	-1.0	-140.0
32.5 - 37.5	30	35.0	0.0	0.0
37.5 - 42.5	40	40.0	1.0	40.0
42.5 - 47.5	60	45.0	2.0	120.0
47.5 - 52.5	10	50.0	3.0	30.0
52.5 - 57.5	20	55.0	4.0	80.0
Total	640			-810.0

We know that $\bar{x} = A + \frac{\sum f_i d_i}{N} \times c$

Choose $A = 35$

$\therefore \quad d_i = \frac{x_i - A}{c} = \frac{x_i - 35}{5}$

$\therefore \quad \bar{x} = A + \frac{\sum f_i d_i}{N} \times c$

(i.e) $\bar{x} = 35 + \frac{(-810)}{640} \times 5$

(i.e) $\bar{x} = 35 - 6.328$

(i.e) $\bar{x} = 28.672$

Exapmle 1. 5 :

Given the following frequency distribution, calculate arithmetic mean.

Interval of marks	No. of students
0 – 10	25
10 – 20	5
20 – 30	15
30 – 40	5
40 – 50	15

Solution :

We know that $\bar{x} = A + \frac{\sum f_i d_i}{N}$

Daily wages in (Rs.)	No. of workers (*f*)	*mid-value* (x_i)	$d_i = \frac{x_i - 15}{10}$	$f_i x_i$
0 - 10	25	5.0	-1.0	-25.0
10 - 20	15	15.0	0.0	0.0
20 - 30	15	25.0	1.0	15.0
30 - 40	05	35.0	2.0	10.0
40 - 50	15	45.0	3.0	45.0
Total	75			45.0

We know that $\bar{x} = A + \frac{\sum f_i d_i}{N} \times c$

Choose $A = 15$

$\therefore \quad d_i = \frac{x_i - A}{c} = \frac{x_i - 10}{10}$

Thus $\bar{x} = A + \frac{\sum f_i d_i}{N} \times c$

Space for Hint

Space for Hint

(i.e) $\overline{x} = 15 + \frac{45}{75} \times 10$

(i.e) $\overline{x} = 15 + 6$

(i.e) $\overline{x} = 21$.

Exapmle 1. 6 :

The mean mark of 20 students in subject is 50. By mistake, marks of two students were entered as 65 and 67 instead of 56 and 76. Find the correct mean mark of the 20 students.

Solution : Given that mean marks of 20 students = 50

(i.e) $\frac{\sum x_i}{n} = 50$

(i.e) $\frac{\sum x_i}{20} = 50$

(i.e) $\sum x_i = 20 \times 50$

(i.e) $\sum x_i = 1000$

(i.e) incorrect $\sum x_i = 1000$

Now correct $\sum x_i$

= incorrect $\sum x_i$ − (sum of incorrect marks) + (sum of correct marks)

$= 1000 - (65 + 67) + (56 + 76)$

$= 1000 - 132 + 132$

$= 1000$

$\therefore$ correct mean $= \frac{\sum x_i}{n}$

$= \frac{1000}{20}$

$= 50$

Hence the correct mean is 50.

Theorem 1. 1

The algebraic sum of the deviation of a set of n values from their arithmetic mean is zero.

Space for Hint

Proof : Let $x_1, x_2, x_3, \cdots, x_n$ be the values having frequencies $f_1, f_2, f_3, \cdots, f_n$ respectively.

Then $\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$ -- (1.1)

Now the deviation of x_i from $\bar{x}$ is $x_i - \bar{x}$

Let $d_i = x_i - \bar{x}$

Thus the sum of deviations of $x_1, x_2, x_3, \cdots, x_n$ from $\bar{x}$

$= \sum f_i d_i$

$= \sum f_i (x_i - \bar{x})$

$= \sum f_i x_i - \sum f_i \bar{x}$

$= \sum f_i x_i - \bar{x} \sum f_i$

$= \sum f_i x_i - \frac{\sum f_i x_i}{\sum f_i} \sum f_i$ (from (1.1))

$= \sum f_i x_i - \sum f_i x_i$

$= 0$

Hence the algebraic sum of the deviation of a set of *n* values from their arithmetic mean is zero.

This proves the theorem.

Theorem 1. 2

The sum of the squares of the deviations of a set of *n* values is minimum when the deviations are taken from their mean.

Proof : Let $x_1, x_2, x_3, \cdots, x_n$ be the values having frequencies $f_1, f_2, f_3, \cdots, f_n$ respectively.

Then $\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$

Now the sum of square of deviations of x_i from an arbitrary value A is given by $y = \sum f_i (x_i - A)^2$

Now we shall find that value of A so that *y* is minimum.

For that differentiate *y* twice with respect to A, we have,

Space for Hint

$$\frac{dy}{dA} = \Sigma\left(f_i 2(x_i - A)(-1)\right)$$

and $$\frac{d^2y}{dA^2} = \Sigma f_i 2(-1)(-1)$$

$$= 2\Sigma f_i$$

$$= 2\Sigma f_i$$

$$= 2N > 0$$

For maximum or minimum y, put $\frac{dy}{dA} = 0$

(i.e) $\Sigma\left(f_i 2(x_i - A)(-1)\right) = 0$

(i.e) $\Sigma f_i x_i - A\Sigma f_i = 0$

(i.e) $-A\Sigma f_i = -\Sigma f_i x_i$

(i.e) $A = \frac{\Sigma f_i x_i}{\Sigma f_i}$

(i.e) $A = \overline{x}$

When $A = \overline{x}$, then $\frac{d^2y}{dA^2} = 2N > 0$

(i.e) y is minimum when $A = \overline{x}$.

Thus the sum of the squares of the deviations of a set of n values is minimum when the deviations are taken from their mean.

Theorem 1. 3 :

If $x_1, x_2, x_3, \cdots, x_k$ are arithmetic means of $n_1, n_2, n_3, \cdots, n_k$ observations, then the arithmetic mean of combined set of observation is $\overline{x} = \frac{n_1x_1 + n_2x_2 + n_3x_3 + \cdots + n_kx_k}{n_1 + n_2 + n_3 + \cdots + n_k}$.

Proof : Let $x_1, x_2, x_3, \cdots, x_k$ be arithmetic means of $n_1, n_2, n_3, \cdots, n_k$ observations.

(i.e) $\overline{x}_1 = \frac{\Sigma x_i}{n_1}$

$\Rightarrow \Sigma x_i = n_1\overline{x}_1$

$\Rightarrow$ sum of all n_1 observation in the first set $= n_1\overline{x}_1$

Similarly, sum of all n_2 observation in the second set = $n_2\bar{x}_2$,

sum of all n_3 observation in the first set = $n_3\bar{x}_3$

$\vdots \quad \vdots \quad \vdots \quad \vdots$

and sum of all n_k observation in the first set = $n_k\bar{x}_k$

Thus the sum of all $n_1 + n_2 + n_3 + \cdots + n_k$ observations

$$= n_1x_1 + n_2x_2 + n_3x_3 + \cdots + n_kx_k$$

Hence the arithmetic mean of the combined set = $\bar{x}$

$$= \frac{n_1x_1 + n_2x_2 + n_3x_3 + \cdots + n_kx_k}{n_1 + n_2 + n_3 + \cdots + n_k}$$

Exapmle 1. 7 :

The mean marks got by 300 students in Statistics was 45. The mean of the top 100 of them was found to be 70 and the mean of the last 100 was known to be 20. Find the mean of the remaining 100 students.

Proof :

Let $\bar{x}_1, \bar{x}_2$ be the mean of the top and last 100 students.

Let $\bar{x}_3$ be the mean of the remaining 100 students.

Given that $n_1 = 100,\ \bar{x}_1 = 70$

$n_2 = 100,\ \bar{x}_2 = 20$

$n_3 = 100,\ \bar{x}_3 = ?$

We know that $\bar{x} = \dfrac{n_1x_1 + n_2x_2 + n_3x_3}{n_1 + n_2 + n_3}$

(i.e) $45 = \dfrac{100 \times 70 + 100 \times 20 + 100 \cdot \bar{x}_3}{100 + 100 + 100}$

(i.e) $7000 + 2000 + 100 \cdot \bar{x}_3 = 45(300)$

(i.e) $9000 + 100 \cdot \bar{x}_3 = 13500$

(i.e) $100 \cdot \bar{x}_3 = 4500$

(i.e) $\bar{x}_3 = 45$

Thus the mean of the remaining 100 students is 45.

Space for Hint

Space for Hint

Exapmle 1. 8 :

Find the missing frequencies to the following distribution whose mean is 56.47.

Wages (in Rs.)	No. of workers
45	5
50	48
55	?
60	30
65	?
70	8
75	6
Total	150

Solution :

Let x and y be the missing frequencies.

We know that $\bar{x} = A + \frac{\sum f_i d_i}{\sum f_i} \times c$

Wages (in Rs.)	No. of workers	$d = \frac{x_i - 60}{5}$	fd
45	5	-3	-15
50	48	-2	-96
55	x	-1	-x
60	30	0	0
65	y	1	y
70	8	2	16
75	6	3	18
Total	150		$y-x-77$

Now $\sum f_i = 150$

(i.e) $97 + x + y = 150$

(i.e) $x + y = 53$ -- (1.2)

Given that $\bar{x} = 56.47$

(i.e) $A + \frac{\sum f_i d_i}{\sum f_i} \times c = 56.47$

(i.e) $60 + \frac{y - x - 77}{150} \times 5 = 56.47$

Space for Hint

(i.e) $\frac{y-x-77}{150} \times 5 = -3.53$

(i.e) $y-x = 28.9$ -- (1.3)

Solving (1.2) and (1.3), we get $x=12$ and $y=41$.

Thus the missing frequencies are 12, 41 respectively.

Exapmle 1. 9 :

The mean yearly salary of employees of a company was Rs. 24000. The mean yearly salaries of male and female employees were Rs. 25000 and Rs. 19000 respectively. Find out the percentages of male and female employees in the company.

Solution :

Let the number of male and female employees be n_1 and n_2 respectively.

Let $\bar{x}_1$ and $\bar{x}_2$ be the mean salaries of male and female employees respectively.

Let $\bar{x}$ be the mean salary of all the workers in the factory.

Given that $n_1+n_2 = 100$ -- (1.4)

Now $n_1 = ?$, $n_2 = ?$

$\bar{x}_1 = 25000$. $\bar{x}_2 = 19000$ and $\bar{x} = 24000$

We know that $\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1+n_2}$

(i.e) $24000 = \frac{n_1(25000)+(100-n_2)(19000)}{100}$

(i.e) $25n_1 + 1900 - 19n_1 = 2400$

(i.e) $n_1 = 83.33$

Again (1.4) $\Rightarrow$ $83.33 + n_2 = 100$

(i.e) $n_2 = 16.67$

Thus the percentage of male employees = 83.33%

and the percentage of female employees = 16.67%

Space for Hint

Exapmle 1. 10 :

Prove that the mean of the values $1,2,3,\cdots,n$ having frequencies $1^2,2^2,3^2,\cdots,n^2$ respectively is $\frac{3}{2}\left(\frac{n^2+n}{2n+1}\right)$.

Proof :

Given that $1,2,3,\cdots,n$ having frequencies $1^2,2^2,3^2,\cdots,n^2$

$$\therefore\ \bar{x} = \frac{1\cdot1^2+2\cdot2^2+3\cdot3^2+\cdots+n\cdot n^2}{1^2+2^2+3^2+\cdots+n^2}$$

$$= \frac{1^3+2^3+3^3+\cdots+n^3}{1^2+2^2+3^2+\cdots+n^2}$$

$$= \frac{\left(\frac{n(n+1)}{2}\right)^2}{\frac{n(n+1)(2n+1)}{6}}$$

$$= = \frac{6}{4}\frac{n^2(n+1)^2}{n(n+1)(2n+1)}$$

$$= \frac{3}{2}\left(\frac{n^2+n}{2n+1}\right)$$

This proves the problem.

Merits and demerits of arithmetic mean

Merits :

- It is easily defined.
- It is easy to understand and easy to calculate.
- It is based on each and every observation of the data.
- It is capable of further mathematical treatment.
- It is affected by fluctuation of sampling.

Demerits :

- It cannot be determined nu inspection nor can it be located graphically.
- It cannot be used in case of open end classes.
- It is very much affected by extreme values.
- It depends on all values of the observation.

Check Your Progress :

(1) Find the mean marks of the students from the following table.

Marks	No. of students	Marks	No. of students
0 & above	80	60 & above	28
10 & above	77	70 & above	16
20 & above	72	80 & above	10
30 & above	65	90 & above	8
40 & above	55	100 & above	0
50 & above	43		

(Answer : Mean = 51.75)

(2) Show that the arithmetic mean of the first n natural numbers is $\frac{1}{2}(n+1)$.

(3) Prove that the weighted arithmetic mean of first n natural numbers whose weight are equal to the corresponding numbers is equal to $\frac{1}{3}(2n+1)$.

1.1.2 Weighted arithmetic mean

Let $x_1, x_2, x_3, \cdots, x_n$ be n numbers. Suppose with each x_i there is associated weight w_i. Then the weighted arithmetic mean or weighted average of $x_1, x_2, x_3, \cdots, x_n$ is defined by $\bar{x}_w = \frac{w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_nx_n}{w_1 + w_2 + w_3 + \cdots w_n}$.

(i.e) $\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$

Exapmle 1. 11 :

Calculate (i) mean price and (ii) weighted mean price of the following food articles from the following table :

Food articles	Quantity (in kgs)	Price per kg
Rice	25	24.0
Wheat	10	12.5
Sugar	5	13.5
Oil	3.5	98.5
Flour	4.5	23.2
Ghee	1	67.0
Onion	10	24.0

Space for Hint

Space for Hint

Solution :

Food articles	Quantity (in kgs) w_i	Price per kg x_i	$w_i x_i$
Rice	25	24.0	600
Wheat	10	12.5	125
Sugar	5	13.5	67.5
Oil	3.5	98.5	344.75
Flour	4.5	23.2	104.4
Ghee	1	67.0	67
Onion	10	24.0	240
Total	59	262.7	1548.65

$$\text{Mean price} = \frac{\sum x}{n}$$

$$= \frac{262.7}{7}$$

$$= \text{Rs. } 37.53$$

and weighted arithmetic mean $= \bar{x}_w = \dfrac{\sum w_i x_i}{\sum w_i}$

$$= \frac{1548.65}{59}$$

$$= \text{Rs. } 26.25$$

1.1. 2 Median

Median is one of the partitioned values. Here the partitioned value, we mean that the variate which divide the total frequency into a number of equal parts. Median is defined as the value of the variate which divides the total frequency into two equal parts.

(i.e) median is the middle value of the variate.

Case (i) : ***Ungrouped data*** : In the case of ungrouped data the median is obtained from the following formula :

$$\text{Median} = \begin{cases} \left(\dfrac{n+1}{2}\right)^{th} \textit{ item if } n \textit{ is odd} \\ \dfrac{1}{2}\left(\left(\dfrac{n}{2}\right)^{th} \textit{item} + \left(\dfrac{n}{2}+1\right)^{th} \textit{item}\right) \textit{ if } n \textit{ is even} \end{cases}$$

Case(ii) : In the case of ***grouped frequency distribution***, median can be obtained from the following formula.

$$\text{Median} = l + \frac{\frac{N}{2} - f_1}{f_2} \times c$$

where l = lower limit of the median class,

f_1 = cumulative frequency of the median class

f_2 = frequency of the median class

c = length of the median class.

Exapmle 1. 12 :

Find the median to the following data :

Marks	No. of students
5 - 10	5
10 - 15	6
15 - 20	15
20 - 25	10
25 - 30	5
30 - 35	4
35 - 40	2
40 - 45	2

Solution :

We know that $\text{Median} = l + \frac{\frac{N}{2} - f_1}{f_2} \times c$

where l = lower limit of the median class,

f_1 = cumulative frequency of the median class

f_2 = frequency of the median class

c = length of the median class.

Space for Hint

Space for Hint

Now

Marks	No. of Students	Cumulative Frequency
5 - 10	5	5
10 - 15	6	11
15 - 20	15	26
20 - 25	10	36
25 - 30	5	41
30 - 35	4	45
35 - 40	2	47
40 - 45	2	49

Here $N = 49$

$\therefore \ \frac{N}{2} = 24.5$

Thus the median class is **15 - 20**

Hence $l = 15, \ f_1 = 11, \ f_2 = 15, \ c = 5$

$$\text{Thus Median} = l + \frac{\frac{N}{2} - f_1}{f_2} \times c$$

$$\text{Median} = 15 + \frac{\frac{49}{2} - 11}{15} \times 5$$

(i.e) median = 19.5

Example 1. 1 :

An incomplete distribution is given below :

Class	Frequency
10 - 20	12
20 - 30	30
30 - 40	?
40 - 50	65
50 - 60	?
60 - 70	25
70 - 80	18
Total	229

Given that the median value is 46. Determine the missing frequencies.

Solution :

Given that

Class	Frequency
10-20	12
20-30	30
30-40	?
40-50	65
50-60	?
60-70	25
70-80	18
Total	229

Let the frequency of the class interval $30-40$ be a and that of class inter $50-60$ be b.

Space for Hint

Space for Hint

Class	Frequency	Cumulative frequency
10 - 20	12	12
20 - 30	30	42
30 - 40	a	$42 + a$
40 - 50	65	$107 + a$
50 - 60	b	$107 + a + b$
60 - 70	25	$132 + a + b$
70 - 80	18	$150 + a + b$
Total	229	

Now $\sum f_i = 229$

(i.e.) $150 + a + b = 229$

(i.e.) $a + b = 79$ -- (1.5)

Given that the median = 46

$\therefore$ the median class is $40 - 50$.

Thus $l = 40$, $f_1 = 42 + a$, $f_2 = 65$, $c = 10$, $N = 229$.

Now median = 46

(i.e.) $l + \dfrac{\frac{N}{2} - f_1}{f_2} \times c = 46$

(i.e.) $40 + \dfrac{\frac{229}{2} - (42 + a)}{65} \times 10 = 46$

(i.e.) $\dfrac{114.5 - 42 - a}{65} \times 10 = 6$

(i.e.) $\dfrac{72.5 - a}{65} = 0.6$

(i.e.) $72.5 - a = 39$

(i.e.) $a = 33.5$

(i.e.) $a \approx 34$

Put $a = 34$ in (1.5) we get $b = 45$.

Thus the missing frequencies of the intervals 30 – 40 and 50 – 60 are 34, 45 respectively.

Space for Hint

1.1. 3 Deciles and Percentiles

Deciles : The values of the variates for which the cumulative frequencies are $\frac{iN}{10}$, $(i = 1, 2, 3, \cdots, 9)$ called deciles.

Note : 1) i^{th} decile is denoted by D_i

2) 5^{th} decile is median of the distribution.

The formula to find i^{th} decile is given by

$$D_i = l + \frac{\frac{iN}{10} - f_1}{f_2} \times c$$

where l = lower limit of the i^{th} decile class,

f_1 = cumulative frequency of the i^{th} decile class

f_2 = frequency of the i^{th} decile class

c = length of the i^{th} decile class.

Percentile : The values of the variates for which the cumulative frequencies are $\frac{iN}{100}$, $(i = 1, 2, 3, \cdots, 99)$ called percentiles.

Note : 1) i^{th} percentiles is denoted by P_i

2) 50^{th} percentiles is median of the distribution.

The formula to find i^{th} percentile is given by

$$P_i = l + \frac{\frac{iN}{100} - f_1}{f_2} \times c$$

where l = lower limit of the i^{th} percentile class,

f_1 = cumulative frequency of the i^{th} percentile class

f_2 = frequency of the i^{th} percentile class

c = length of the i^{th} percentile class.

Space for Hint

Example 1. 2 :

Find the median, 8^{th} decile and 56^{th} percentile for the following data.

Monthly wages	Frequency	Monthly wages	Frequency
1 - 2.99	6	9 - 10.99	21
3 - 4.99	53	11 - 12.99	16
5 - 6.99	85	13 - 14.99	4
7 - 8.99	56	15 - 16.99	4

Solution :

Step 1 : To find the eighth decile.

Given that

Monthly wages	Frequency	Cumulative frequency
0.995 – 2.995	6	6
2.995 – 4.995	53	59
4.995 – 6.995	85	144
6.995 – 8.995	56	200
8.995 – 10.995	21	221
10.995 – 12.995	16	237
12.995 – 14.995	4	241
14.995 – 16.995	4	245

We know that the formula for finding i^{th} decile is given by

$$D_i = l + \frac{\frac{iN}{10} - f_1}{f_2} \times c$$

where l = lower limit of the i^{th} decile class,

f_1 = cumulative frequency of the i^{th} decile class

f_2 = frequency of the i^{th} decile class

c = length of the i^{th} decile class.

Here N = 245

$\therefore \ \frac{8\times N}{10} = \frac{8\times 245}{10}$

$= 196$

Thus the 8^{th} decile class is 6.995 – 8.995

Hence $l = 6.995,\ f_1 = 144,\ f_2 = 56,\ c = 2$

Now $D_8 = l + \frac{\frac{iN}{10} - f_1}{f_2} \times c$

$= 6.995 + \frac{196-144}{56} \times 2$

$= 6.995 + 1.8571$

$= 8.8521$

Thus the 8^{th} decile = 8.8521.

Step 2 : To find p_{56}.

Monthly wages	Frequency	Cumulative frequency
0.995 – 2.995	6	6
2.995 – 4.995	53	59
4.995 – 6.995	85	144
6.995 – 8.995	56	200
8.995 – 10.995	21	221
10.995 – 12.995	16	237
12.995 – 14.995	4	241
14.995 – 16.995	4	245

Space for Hint

Space for Hint

We know that The formula to find i^{th} percentile is given by

$$P_i = l + \frac{\frac{iN}{100} - f_1}{f_2} \times c$$

where l = lower limit of the i^{th} percentile class,

f_1 = cumulative frequency of the i^{th} percentile class

f_2 = frequency of the i^{th} percentile class

c = length of the i^{th} percentile class.

Thus $P_{56} = l + \frac{\frac{56N}{100} - f_1}{f_2} \times c$

Here N = 245

$\therefore \frac{56 \times N}{100} = \frac{56 \times 245}{100}$

$= 122.5$

Thus the P_{56} class is 4.995 – 6.995

Hence $l = 4.995,\ f_1 = 59,\ f_2 = 85,\ c = 2$

Now $P_{56} = l + \frac{\frac{56N}{100} - f_1}{f_2} \times c$

$= 4.995 + \frac{137.5 - 59}{85} \times 2$

$= 4.995 + 1.84471$

$= 6.8421$

Thus the 8^{th} decile = 6.8421.

Example 1. 3

From the following data calculate the percentage of workers getting wages

(i) more than Random sample. 44 and

(ii) between Random sample. 22 and Random sample. 58.

Space for Hint

Wages in Rs.	No. of workers
0 – 10	20
10 – 20	45
20 – 30	85
30 – 40	160

Wages in Rs.	No. of workers
40 – 50	70
50 – 60	55
60 – 70	35
70 – 80	30

Solution :

Number of workers getting more than Rs. 44 as wage.

$$= \frac{50-44}{10} \times 70 + 55 + 35 + 30$$

$= 162$

Therefore percentage of workers getting more than Rs. 44 as wages

$$= \frac{162}{500} \times 100$$

$= 32.4\%$

Again the number workers getting between Rs. 32 and Rs. 58 as wage

$$= \frac{30-22}{10} \times 85 + 160 + 70 + \frac{58-50}{10} \times 55$$

$= 342$

Therefore percentage of workers getting between Rs. 22 and Rs. 58 as wages

$$= \frac{342}{500} \times 100$$

$= 68.4\ \%$

MERITS AND DEMERITS

Merits

- It is rigidly defined.
- It is easy to understand and easy to calculate for a non-mathematical person.

Demerits

- It is not suitable for further mathematical treatment.
- It is more affected by sampling fluctuations than arithmetic mean and therefore it is less reliable.

Space for Hint

Check your progress :

Find the (i) mean, (ii) median, (iii) first quartile, (iv) third quartile, (v) 9^{th} decile and (vi) 19^{th} percentile for the following frequency distribution.

Class	Frequency	Class	Frequency
11 – 15	8	36 – 40	41
16 – 20	15	41 – 45	28
21 – 25	39	46 – 50	16
26 – 30	47	51 – 55	4
31 – 35	52	Total	250

(Answer :

(i) mean = 32.18, (ii) median = 32.04, (iii) first quartile = 25.55,

(iv) third quartile = 38.75, (v) 9^{th} decile = 41.61 and

(vi) 19^{th} percentile = 23.64

1.1. 4 Mode

Definition : The value of the variable which occurs maximum number of times is called **mode** of the distribution.

In the case of frequency distribution, the mode can be calculated from

$$\text{mode} = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2}$$

where l = lower limit of the modal class,

f_1 = cumulative frequency of the modal class

f_2 = frequency of the modal class

c = length of the modal class.

Space for Hint

Note :

(1) A frequency distribution may have more than one mode called as multi-modal distribution. If there is only one mode then the distribution is called unimodal distribution.

(2) The relationship between mean, median and mode is called an **empirical realation** and it is given by $Mode = 3Median - 2Mean$.

Example 1. 4 :

Calculate the mode of the following distribution

6, 8, 2, 5, 9, 5, 6, 5, 2, 3

Solution :

Given that 6, 8, 2, 5, 9, 5, 6, 5, 2, 3

Since 5 is repeated maximum number of times and therefore 5 is the mode of the distribution.

Example 1. 5

Find the mode of the following frequency distribution

Class	:	20 –24	25-29	30-34	35-39	40-44	45-49	50-54	55-59
Frequency	:	3	5	10	20	12	6	3	1

Solution :

Given that

Class	:	20 –24	25-29	30-34	35-39	40-44	45-49	50-54	55-59
Frequency	:	3	5	10	20	12	6	3	1

Step 1 : First we shall find the modal class.

Space for Hint

class	frequency	II	III	IV	V	VI
20-24	3					
		8				
25-29	5		18			
		15		35		38
30-34	10				42	
		30				
35-39	**20**					
		32				
40-44	12		38			
		18				
45-49	6			21		
		9			10	22
50-54	3					
		4				
55-59	1					

Step 2 :

Analysis table

class	Number of times which occurs as maximum frequency	Frequency
20-24	I	1
25-29	II	2
30-34	IIII	4
35-39	~~IIII~~ I	6
40-44	II	2
45-49	I	1
50-54	0	0
55-59	0	0

From the above table it is clear that 35 – 39 is the modal class.

Thus $l = 34.5,\ f_0 = 10,\ f_1 = 20,\ f_2 = 12$.

We know that

$$\text{mode} = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2}$$

where l = lower limit of the modal class,

f_1 = cumulative frequency of the modal class

f_2 = frequency of the modal class

c = length of the modal class.

$$\text{Hence mode} = 34.5 + \frac{20-10}{2\times 20-(10+12)} \times 5$$

$$= 34.5 + \frac{10}{40-22} \times 5$$

$$= 34.5 + 2.778$$

$$= 37.278$$

Thus the mode of the distribution is 37.278

Example 1. 6 :

Find the mode to the following data :

Class	Frequency
0 – 9	6
10 – 19	29
20 – 29	87
30 – 39	181
40 – 49	247

Class	Frequency
50 – 59	263
60 – 69	133
70 – 79	43
80 – 89	9
90 – 99	2

Solution :

Step 1 : First we shall find the modal class.

Space for Hint

Space for Hint

class	frequency	II	III	IV	V	VI
0-9	6					
		35				
10-19	29			122		
			116			
20 - 29	87				297	
		268				
30-39	181					515
			428			
40-49	247			691		
		510				
50-59	263				643	
			396			
60-69	133					439
		176				
70-79	43			185		
			52			
80-89	9				54	
		11				
90-99	2					

Step 2 :

Analysis table

class	Number of times which occurs as maximum frequency	Frequency
30 – 39	III	3
40 – 49	~~IIII~~	5
50 – 59	IIII	4
60 – 69	0	0
70 – 79	0	0

From the above table it is clear that 40 – 49 is the modal class.

Thus $l = 39.5,\ f_0 = 181,\ f_1 = 247,\ f_2 = 263,\ C = 10$.

We know that

$$\text{mode} = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2}$$

where l = lower limit of the modal class,

f_1 = cumulative frequency of the modal class

f_2 = frequency of the modal class

c = length of the modal class.

$$\text{Hence mode} = 39.5 + \frac{247-181}{2\times247-(181+263)}\times10$$

$$= 39.5 + \frac{66}{50}\times10$$

$$= 39.5 + 13.2$$

$$= 52.7$$

Thus the mode of the distribution is 52.7.

1.1. 5 Geometric Mean

Definition : The geometric mean of the individual observations $x_1, x_2, x_3, \cdots, x_n$ is defined as G.M. = $(x_1 \cdot x_2 \cdot x_3 \cdots x_n)^{1/n}$.

Note : Geometric mean can be calculated from

$$\text{G.M} = Anti\log\left(\frac{1}{n}\log\left(\sum x_i\right)\right)$$

Definition : The Geometric mean of the frequency distribution is defined as

$G.M. = \left(x_1^{f_1} \cdot x_2^{f_2} \cdot x_3^{f_3} \cdots x_n^{f_n}\right)^{1/N}$ where $N = \sum x_i$

Note : Geometric mean can be calculated from

$$\text{G.M} = Anti\log\left(\frac{1}{n}\log\left(\sum x_i\right)\right).$$

Space for Hint

Space for Hint

Example 1. 7 :

Find the geometric mean of 32, 35, 36, 37, 39, 41, 43.

x	$\log x$
32	1.5051
35	1.5441
36	1.5563
37	1.5682
39	1.5911
41	1.6128
43	1.6335
Total	11.0110

We know that G.M = $Anti\log\left(\frac{1}{n}\log\left(\sum x_i\right)\right)$

$$\text{(i.e.) G.M.} = Anti\log\left(\frac{11.0110}{7}\right)$$

$$= Anti\log(1.5730)$$

$$= Anti\log(1.5730)$$

$$= 37.4115$$

Thus the geometric mean of the distribution is 37.4115.

Example 1. 8 :

Find the geometric mean for the following distribution.

Class	Frequency
0 – 9	32
10 – 19	65
20 – 29	100
30 – 39	184
40 – 49	288

Class	Frequency
50 – 59	167
60 – 69	98
70 – 79	46
80 – 89	20
Total	1000

Solution :

We know that G.M = $Anti\log\left(\frac{1}{n}\log\left(\sum x_i\right)\right)$.

Thus

Class	Frequency	mid-value	$\log x$	$f_i \log x_i$
0 – 9	32	4.5	0.6532	20.9028
10 – 19	65	14.5	1.1614	75.48892
20 – 29	100	24.5	1.3892	138.9166
30 – 39	184	34.5	1.5378	282.9587
40 – 49	288	44.5	1.6484	474.7277
50 – 59	167	54.5	1.7364	289.9782
60 – 69	98	64.5	1.8096	177.3369
70 – 79	46	74.5	1.8722	86.1192
80 – 89	20	84.5	1.9269	38.5371
Total	1000			1584.9661

Thus G.M = $Anti\log\left(\frac{1}{n}\log\left(\sum x_i\right)\right)$.

$= Anti\log\left(\frac{1584.9661}{1000}\right)$

$= Anti\log(1.5850)$

= 38.4562.

Thus the geometric mean is 38.4562

MERITS & DEMERITS OF GEOMETRIC MEAN

MERITS :

It is based on all observations.

It is rigidly defined.

It is capable of further algebraic treatment.

It is less affected by the extreme values.

It is useful in studying economic and Social data.

Space for Hint

Space for Hint

DEMERITS :

(1) It is difficult to understand.

(2) Non mathematics persons feel difficult to calculate geometric mean.

(3) The geometric mean cannot be computed if any item in the series is negative or zero.

(4) It has restricted application.

USES OF GEOMETRIC MEAN

Geometric mean is especially useful in the following cases:-

The Geometric mean is used to find the average, percent increase in sales production, population or other economic or business series.

Geometric mean is important in the construction of index numbers.

In economic and social science, where we want to give more weight to smaller items and smaller weight to large items, geometric mean is appropriate.

Note :

Geometric Mean defined as the n^{th} root of the product of n items or values. If there are two items, we take the square root: and so on. The geometric mean is never larger than the arithmetic mean.

Check your progress :

Find the geometric mean to following frequency distribution

Marks	No. of students
130	4
135	5
140	7
145	7
146	4
148	6
150	2
156	2

Space for Hint

1.1.6 Harmonic Mean

Definition : The harmonic mean of the individual observations $x_1, x_2, x_3, \cdots, x_n$ is defined as

$$H.M. = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}+\cdots+\frac{1}{x_n}}.$$

In the case of grouped frequency we can use the following harmonic mean formula.

$$H.M. = \frac{N}{\frac{f_1}{x_1}+\frac{f_2}{x_2}+\frac{f_3}{x_3}+\cdots+\frac{f_n}{x_n}} \text{ where } N = \sum x_i.$$

Example 1. 9 :

Find harmonic mean to the following frequency distribution

Marks	No. of students
130	4
135	5
140	7
145	7
146	4
148	6
150	2
156	2

Solution :

We know that $H.M. = \frac{N}{\frac{f_1}{x_1}+\frac{f_2}{x_2}+\frac{f_3}{x_3}+\cdots+\frac{f_n}{x_n}}$ where $N = \sum x_i$.

Space for Hint

Now

Marks	No. of students	f_i/x_i
130	4	0.0308
135	5	0.0370
140	7	0.0500
145	7	0.0483
146	4	0.0274
148	6	0.0405
150	2	0.0133
156	2	0.0128
Total	37	0.260174

Thus harmonic mean = $\dfrac{N}{\dfrac{f_1}{x_1}+\dfrac{f_2}{x_2}+\dfrac{f_3}{x_3}+\cdots+\dfrac{f_n}{x_n}}$

$= \dfrac{37}{0.2602}$

$= 142.1983$

Hence harmonic mean = 142.1983

Example 1. 10 :

Find the harmonic mean for the following data.

x	:	2	3	4	5	6
f	:	5	7	11	9	8

Solution :

We know that harmonic mean = $\dfrac{N}{\dfrac{f_1}{x_1}+\dfrac{f_2}{x_2}+\dfrac{f_3}{x_3}+\cdots+\dfrac{f_n}{x_n}}$

Now

x	f	$\frac{f}{x}$
2	5	2.5000
3	7	2.3333
4	11	2.7500
5	9	1.8000
6	8	1.3333
total	40	10.7167

Thus harmonic mean = $\dfrac{N}{\dfrac{f_1}{x_1}+\dfrac{f_2}{x_2}+\dfrac{f_3}{x_3}+\cdots+\dfrac{f_n}{x_n}}$

$= \dfrac{10.7167}{40}$

$= 3.7325$

Hence harmonic mean = 3.7325

MERITS & DEMERITS OF HARMONIC MEAN

MERITS :

It is based on all observations.

It is rigidly defined.

It is capable of further algebraic treatment.

It the most suitable average when it is desired to give greater weightage to smaller observations and less weightage to the largest ones.

DEMERITS :

It is not easily to understand.

Non mathematics persons feel difficult to calculate geometric mean.

It is a summary figure and may not be the actual item in the series.

Gives greater importance to small values and is useful only when small items have to be given greater weightage.

Space for Hint

Space for Hint

1. 2 Measures of Dispersions

In the previous sections we have discussed measures of central tendencies and they give an idea of the concentration of the data about the central part of the distribution. however these measures are not sufficient to study the complete nature of the distribution. The measures of dispersions are supported and supplemented by some other measures. One of such measures is measure of distributions.

Definition : The amount of scattering of the individual observations from the measure of central tendency is called ***dispersion*** of the distribution.

The commonly used dispersions are

(1) Range

(2) Mean deviation

(3) Quartile deviation

(4) Standard deviation

1.2. 1 Range

Range is the simplest dispersion and is defined as the difference between maximum and minimum values of the variate.

(i.e.) Range = L – S

where L = largest value of the distribution

and S = smallest value of the distribution.

The coefficient of range = $\frac{L-S}{L+S}$

Example 1. 11 :

Find the range and its coefficient for the following data

100, 120, 140, 120, 180, 175, 185, 130, 200, 150.

Solution :

Given that 100, 120, 140, 120, 180, 175, 185, 130, 200, 150.

Here L = 200 and S = 100

$\therefore$ range = $L-S$ = 200 – 100 = 100

and the coefficient of range = $\frac{L-S}{L+S}$

$$= \frac{200-100}{200+100}$$

$$= \frac{100}{300}$$

$$= \frac{1}{3}.$$

1.2. 2 Mean Deviation

The Mean Deviation of a frequency distribution about A is given by

$$\text{M.D. about A} = \frac{1}{N}\sum_{i=1}^{n} f_i \,|\, x_i - A \,| \quad \text{where } N = \sum_{i=1}^{n} f_i .$$

$$\text{The coefficient of mean deviation} = \frac{M.D.}{A}$$

Note : If the value if A is not given then it is taken as either mean or median.

Example 1. 12 :

Find the Mean Deviation about mean from the following data

x	f
10	3
11	12
12	18
13	12
14	3

Solution :

Step 1 : First we shall find mean of the distribution

x	f	fx
10	3	30
11	12	132
12	18	216
13	12	156
14	3	42
Total	48	576

Space for Hint

Space for Hint

We know that $\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$

$= \frac{576}{48}$

$= 12.$

Step 2 : To find the Mean Deviation about mean.

We know that Mean Deviation about mean = $\frac{1}{N}\sum_{i=1}^{n} f_i \mid x_i - \bar{x} \mid$

x	f	$x_i - \bar{x}$	$\mid x_i - \bar{x} \mid$	$f_i \mid x_i - \bar{x} \mid$
10	3	-2	2	6
11	12	-1	1	12
12	18	0	0	0
13	12	1	1	12
14	3	2	2	6
Total	48	0	6	36

Thus the Mean Deviation about mean = $\frac{1}{N}\sum_{i=1}^{n} f_i \mid x_i - \bar{x} \mid$

$= \frac{36}{48}$

$= 0.75$

Example 1. 13 :

Calculate the mean deviation from (i) mean and (ii) median from the following data

size of item	Frequency
3 - 4	3
4 - 5	7
5 - 6	22
6 - 7	60
7 - 8	85
8 - 9	32
9 - 10	8

Space for Hint

Solution :

Step 1 : First we shall find mean of the distribution

size of item	Frequency	mid-value	$d=\frac{x-6.5}{1}$	fd
3 - 4	3	3.5	-3	-9
4 - 5	7	4.5	-2	-14
5 - 6	22	5.5	-1	-22
6 - 7	60	6.5	0	0
7 - 8	85	7.5	1	85
8 - 9	32	8.5	2	64
9 - 10	8	9.5	3	24
Total	217			128

Choose $A=6.5$

We know that $\bar{x}=A+\frac{\sum fd}{N}\times c$

Space for Hint

$$= 6.5 + \frac{128}{217} \times 1$$

$$= 7.0899 .$$

$$= 7.09 .$$

Step 2 : To find the Mean Deviation about mean.

We know that Mean Deviation about mean = $\frac{1}{N}\sum_{i=1}^{n} f_i \mid x_i - \bar{x} \mid$

size of item	Frequency	Mid-value	$x_i - \bar{x}$	$\mid x_i - \bar{x} \mid$	$f_i \mid x_i - \bar{x} \mid$
3 - 4	3	3.5	-3.59	3.59	10.77
4 - 5	7	4.5	-2.59	2.59	18.13
5 - 6	22	5.5	-1.59	1.59	34.98
6 - 7	60	6.5	-0.59	0.59	35.4
7 - 8	85	7.5	0.41	0.41	34.85
8 - 9	32	8.5	1.41	1.41	45.12
9 - 10	8	9.5	2.41	2.41	19.28
Total	217				198.53

Thus the Mean Deviation about mean = $\frac{1}{N}\sum_{i=1}^{n} f_i \mid x_i - \bar{x} \mid$

$$= \frac{198.53}{217}$$

$$= 0.914885$$

Step 3 : Now we shall find the median of the distribution.

We know that median = $l + \frac{\frac{N}{2} - f_1}{f_2} \times c$.

size of item	Frequency	cf
3 - 4	3	3
4 - 5	7	10
5 - 6	22	32
6 - 7	60	92
7 - 8	*85*	*177*
8 - 9	32	209
9 - 10	8	217

Now $\frac{N}{2} = \frac{217}{2} = 108.5$

Therefore the median class is 7 – 8

Thus l = lower limit of the median class = 7

f_1 = cumulative frequency of the median class = 92

f_2 = frequency of the median class = 85

c = length of the median class = 1.

Hence median = $7 + \frac{108.5 - 92}{85} \times 1$

= 7.194118

= 7.19

Step 4 : To find the Mean Deviation about median.

We know that Mean Deviation about mean = $\frac{1}{N}\sum_{i=1}^{n} f_i \,|\, x_i - MD \,|$

Space for Hint

Space for Hint

size of item	Frequency	Mid-value	$x_i - Med$	$\lvert x_i - Med \rvert$	$f_i \lvert x_i - Med \rvert$
3 - 4	3	3.5	-3.69	3.69	11.07
4 - 5	7	4.5	-2.69	2.69	18.83
5 - 6	22	5.5	-1.69	1.69	37.18
6 - 7	60	6.5	-0.69	0.69	41.4
7 - 8	85	7.5	0.31	0.31	26.35
8 - 9	32	8.5	1.31	1.31	41.92
9 - 10	8	9.5	2.31	2.31	18.48
Total	217				195.23

Thus the Mean Deviation about median $= \frac{1}{N}\sum_{i=1}^{n} f_i \lvert x_i - Med \rvert$

$$= \frac{195.23}{217}$$

$$= 0.90$$

Merits and Demerits

Merits :

It is easy to understand

It is rigidly defined

It shows the significance of an average of the distribution

It is less affected by extreme values as compared with standard deviation.

Demerits :

It is not capable of mathematical treatment

It cannot be calculated for distributions with open-end classes.

Space for Hint

1.2. 3 Quartile Deviation

The Quartile Deviation or semi-inter quartile range is defined by

Quartile Deviation = $\frac{1}{2}(Q_3 - Q_1)$

where Q_1 = lower quartile deviation and

and Q_3 = upper quartile deviation

and the coefficient of Quartile Deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$.

Example 1. 14

Calculate the Quartile Deviation to the following data.

x	f
48	3
49	7
50	11
51	14
52	18
53	17
54	13
55	8
56	5
57	4
Total	100

Space for Hint

Solution :

Given that

x	f	cf
48	3	3
49	7	10
50	11	21
51	14	35
52	18	53
53	17	70
54	13	83
55	8	91
56	5	96
57	4	100

We know that Quartile Deviation $= \frac{1}{2}(Q_3 - Q_1)$

Here $Q_1 = \left(\frac{N+1}{4}\right)^{th}$ item

$= 25.25^{th}$ item

$= 51$

Here $Q_3 = \left(\frac{3(N+1)}{4}\right)^{th}$ item

$= 75.75^{th}$ item

$= 54$

Therefore Quartile Deviation $= \frac{1}{2}(Q_3 - Q_1)$

$= \frac{1}{2}(54 - 51)$

$= \frac{3}{2}$

Example 1. 15 :

Find Quartile Deviation and its coefficient to the following distribution.

Income	No. of workers
less than 500	54
500 - 700	100
700 - 900	140
900 - 1100	300
1100 - 1300	230
1300 - 1500	125
above 1500	51

Space for Hint

Solution :

Given that

Income	No. of workers
less than 500	54
500 - 700	100
700 - 900	140
900 - 1100	300
1100 - 1300	230
1300 - 1500	125
above 1500	51

Step 1 : First we shall find lower quartile Q_1

Now

Income	No. of workers	cf
300 - 500	54	54
500 - 700	100	154
700 - 900	***140***	***294***
900 - 1100	300	594
1100 - 1300	230	824
1300 - 1500	125	949
1500 - 1700	51	1000

We know that Lower Quartile = $l + \frac{\frac{N}{4} - f_1}{f_2} \times c$

Space for Hint

where l = lower limit of the Q_1 class,

f_1 = cumulative frequency of just above the Q_1 class

f_2 = frequency of the Q_1 class

c = length of the Q_1 class.

Here N = 1000.

$$\therefore \frac{N}{4} = \frac{1000}{4} = 250$$

Thus $l = 700$, $f_1 = 154$, $f_2 = 140$, $c = 200$.

$$\therefore Q_1 = l + \frac{\frac{N}{4} - f_1}{f_2} \times c$$

$$= 700 + \frac{250 - 154}{140} \times 200$$

$$= 700 + 137.1429$$

$$= 837.1429.$$

Step 2 : First we shall find lower quartile Q_3

Now

Income	No. of workers	cf
300 - 500	54	54
500 - 700	100	154
700 - 900	*140*	*294*
900 - 1100	300	594
1100 - 1300	**230**	**824**
1300 - 1500	125	949
1500 - 1700	51	1000

Space for Hint

We know that Upper Quartile = $l+\dfrac{\frac{3N}{4}-f_1}{f_2}\times c$

where l = lower limit of the Q_3 class,

f_1 = cumulative frequency of just above the Q_3 class

f_2 = frequency of the Q_3 class

c = length of the Q_3 class.

Here N = 1000.

$$\therefore \frac{3N}{4}=\frac{3\times 1000}{4}=750$$

Thus $l=1100$, $f_1=594$, $f_2=230$, $c=200$.

$$\therefore Q_3 = l+\frac{\frac{3N}{4}-f_1}{f_2}\times c$$

$$= 1100+\frac{750-594}{230}\times 200$$

$$= 1100+135.6522$$

$$= 1235.652.$$

Step 3 : First we shall find Quartile Deviation.

Now $Q.D=\frac{1}{2}(Q_3-Q_1)$

$$=\frac{1}{2}(1235.652-847.1429)$$

$$= 194.2515.$$

Step 4 : First we shall find the coefficient of Quartile Deviation.

Now coefficient of Quartile Deviation = $\dfrac{Q_3-Q_1}{Q_3+Q_1}$

$$=\frac{1235.652-847.1429}{1235.652+847.1429}$$

$$= 0.186533.$$

Space for Hint

1.2. 4 Standard Deviation

Definition : Let $x_1, x_2, x_3, \cdots, x_n$ be n observations. Then the standard deviation of these values is defined as $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

Definition : Let $x_1, x_2, x_3, \cdots, x_n$ having the frequencies $f_1, f_2, f_3, \cdots, f_n$ respectively. Then the standard deviation of these values is defined as $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - \bar{x})^2}$ where $N = \sum_{i=1}^{n} f_i$.

Definition : The coefficient of variation of a distribution is defined as $C.V. = \frac{\sigma}{\bar{x}} \times 100$.

Definition : The root mean square deviation of a frequency distribution is defined as $s = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - A)^2}$ where A is given number.

Note 1 : Square of the standard deviation is called variance.

(i.e) variance = σ^2

Note 2 : A variate X is more stable or more consistent or more reliable than a variate Y if $C.V.\ of\ Y > C.V.\ of\ X$.

Example 1. 16 :

Prove that $\sigma^2 = s^2 - d^2$ where $d = \bar{x} - A$.

Proof : Let A be a given number.

Let $d = \bar{x} - A$.

We know that $s = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - A)^2}$

$$\therefore\ s^2 = \frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - A)^2$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - \bar{x} + \bar{x} - A)^2$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i \left[(x_i - \overline{x}) + (\overline{x} - A)\right]^2$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i \left[(x_i - \overline{x})^2 + (\overline{x} - A)^2 + 2(x_i - \overline{x})(\overline{x} - A)\right]$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{x})^2 + \frac{1}{N}\sum_{i=1}^{n} f_i (\overline{x} - A)^2 + 2(\overline{x} - A)\frac{1}{N}\sum_{i-1}^{n} (x_i - \overline{x})$$

$$= \sigma^2 + s^2 + 0$$

$$= \sigma^2 + s^2$$

Hence $\sigma^2 = s^2 - d^2$.

This proves the problem.

Note : From the above example it is clear that standard deviation is minimum possible root mean square deviation.

Example 1. 17 :

Prove that $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{n} f_i x_i\right)^2}$

Proof : We know that $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{x})^2}$

(i.e) $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{x})^2$

(i.e) $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i^2 - 2x_i\overline{x} - \overline{x}^2)$

(i.e) $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^2 - \frac{1}{N}2\overline{x}\sum_{i=1}^{n} f_i x_i + \frac{1}{N}\sum_{i=1}^{n} f_i \overline{x}^2$

(i.e) $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^2 - 2\overline{x}\,\overline{x} + \frac{1}{\not{N}}\not{N}\,\overline{x}^2$

(i.e) $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^2 - \overline{x}^2$

(i.e) $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{n} f_i x_i\right)^2}$

This proves the problem.

Space for Hint

Space for Hint

Example 1. 18 :

Prove that standard deviation is independent of change of origin.

Proof :

Let $x_1, x_2, x_3, \cdots, x_n$ be the values having frequencies $f_1, f_2, f_3, \cdots, f_n$ respectively.

Then $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - \bar{x})^2}$

Suppose if we shift the origin to A then the new values be $u_1, u_2, u_3, \cdots, u_n$.

(i.e) $u_i = x_i - A$ for $i = 1, 2, 3, \cdots, n$.

Now $u_i = x_i - A$

$$\Rightarrow \bar{u} = \frac{1}{N}\sum_{i=1}^{n} f_i\,u_i$$

$$\Rightarrow \bar{u} = \frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - A)$$

$$\Rightarrow \bar{u} = \bar{x} - A$$

$$\text{Thus } \sigma_u = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i\,(u_i - \bar{u})^2}$$

$$= \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i\,\big((x_i - A) - (\bar{x} - A)\big)^2}$$

$$= \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i\,\left(x_i - \cancel{A} - \bar{x} + \cancel{A}\right)^2}$$

$$= \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - \bar{x})^2}$$

$$= \sigma_x$$

$\therefore$ standard deviation is independent on change of origin.

This proves the problem.

Example 1. 19 :

Show that standard deviation is depend on change of scale.

Proof :

Let $x_1, x_2, x_3, \cdots, x_n$ be the values having the frequencies $f_1, f_2, f_3, \cdots, f_n$ respectively.

Suppose we change the scale of x_i to u_i where $u_i = \frac{x_i}{h}$.

Now $u_i = \frac{x_i}{h}$.

$$\therefore \ \bar{u} = \frac{1}{N}\sum_{i=1}^{n} u_i$$

$$= \frac{1}{N}\sum_{i=1}^{n} \frac{x_i}{h}$$

$$= \frac{1}{h}\bar{x}.$$

Hence $\sigma_u = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i \, (u_i - \bar{u})^2}$

$$= \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i \left(\frac{1}{h}x_i - \frac{1}{h}\bar{x}\right)^2}$$

$$= \frac{1}{h}\sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i \, (x_i - \bar{x})^2}$$

$$= \frac{1}{h}\sigma_x$$

(i.e) $\sigma_x = h\sigma_u$.

$\therefore$ standard deviation is depend on change of scale.

This proves the problem.

Example 1. 20 :

Calculate the standard deviation from the following data of income of 10 employees of a firm.

100, 120, 140, 120, 180, 175, 185, 130, 200, 150,

Space for Hint

Space for Hint

Solution :

	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	100	-50	2500
	120	-30	900
	140	-10	100
	120	-30	900
	180	30	900
	175	25	625
	185	35	1225
	130	-20	400
	200	50	2500
	150	0	0
total	1500		10050

Now mean = $\bar{x}$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$= \frac{1}{10}(1500)$$

$$= 150$$

and standard deviation = σ

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \sqrt{\frac{1}{10}(10050)}$$

$$= 31.70173$$

Hence standard deviation of the distribution is 31.70173

Example 1. 21 :

Find the standard deviation of the following height of 100 students.

Height in inches	Number of students
60 - 62	5
63 - 65	18
66 - 68	42
69 - 71	27
72 - 74	8
total	100

Solution :

We know that $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - \bar{x})^2}$

Choose A = 67

$\therefore\ d_i = \frac{x_i - A}{c}$

(i.e) $d_i = \frac{x_i - 67}{3}$.

Height in inches	No. of students	Mid value	$d_i = \frac{x_i - 67}{3}$	$f_i\,d_i$	$f_i\,d_i^2$
60 - 62	5	61	-2	-10	20
63 - 65	18	64	-1	-18	18
66 - 68	42	67	0	0	0
69 - 71	27	70	1	27	27
72 - 74	8	73	2	16	32
total	100	335	0	15	97

We know that standard deviation = σ

$= \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i\,d_i^2 - \left(\frac{1}{N}\sum_{i=1}^{n} f_i\,d_i\right)^2}\,,\times c$

$= \sqrt{\frac{1}{100}(97) - \left(\frac{15}{100}\right)^2}$

= 2.9202

Space for Hint

Space for Hint

Example 1. 22 :

Calculate the mean and standard deviation to the following data :

size of item	frequency	size of item	frequency
3 - 4	3	7 - 8	85
4 - 5	7	8 - 9	32
5 - 6	22	9 - 10	8
6 - 7	60	Total	217

Solution :

We know that mean $\overline{x} = A + \frac{\sum_{i=1}^{n} f_i x_i}{N}$

and $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{n} f_i x_i\right)^2} \times c$

Choose A = 67

$\therefore\ d_i = \frac{x_i - A}{c}$

(i.e) $d_i = \frac{x_i - 67}{3}$

size of item	frequency	Mid value	$d_i = \frac{x_i - 7.5}{1}$	$f_i d_i$	$f_i d_i^2$
3 - 4	3	3.5	-4	-12	48
4 - 5	7	4.5	-3	-21	63
5 - 6	22	5.5	-2	-44	88
6 - 7	60	6.5	-1	-60	60
7 - 8	85	7.5	0	0	0
8 - 9	32	8.5	1	32	32
9 - 10	8	9.5	2	16	32
total	217			-89	323

Now $\overline{x} = A + \frac{\sum_{i=1}^{n} f_i x_i}{N}$

(i.e) $\overline{x} = 7.5 + \frac{(-89)}{217}$

(i.e) $\overline{x} = 7.0899$

We know that $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{n} f_i x_i\right)^2} \times c$

$= \sqrt{\frac{323}{217} - \left(\frac{-89}{217}\right)^2} \times 1$

$= 1.14903$

Example 1. 23 :

The scores of two golf players of 10 rounds each are given below

A	:	58	59	60	54	65	66	52	75	69	52
B	:	84	56	92	65	86	78	44	54	78	68

Who may be considered to be a more consistent player?

Solution :

Step 1 : First we shall find coefficient of variation of A.

x	$d = x - A$	d^2
58	-2	4
59	-1	1
60	0	0
54	-6	36
65	5	25
66	6	36
52	-8	64
75	15	225
69	9	81
52	-8	64
total	10	536

Choose $A = 60$

$\therefore\ d = x - A = x - 60$

We know that $\bar{x} = A + \frac{\sum x_i}{n}$

(i.e) $\bar{x} = 60 + \frac{10}{10}$

(i.e) $\bar{x} = 61.$

and variance $= \sigma^2$

Space for Hint

Space for Hint

$$= \frac{1}{n}\sum d_i^2 - \left(\frac{\sum d_i}{n}\right)^2$$

$$= \frac{1}{10}(536) - \left(\frac{10}{10}\right)^2$$

$$= 52.6$$

$\therefore$ standard deviation = σ_A = 7.253

Thus coefficient of variation for A = CV_A

$$= \frac{\sigma_A}{\overline{x}} \times 100$$

$$= \frac{7.253}{61} \times 100$$

$$= 11.89$$

Step 2 : To find the coefficient of variation for B

y	$d = y - B$	d^2
84	19	361
56	-9	81
92	27	729
65	0	0
86	21	441
78	13	169
44	-21	441
54	-11	121
78	13	169
68	3	9
Total	55	2521

Choose $A = 65$

$\therefore\ d = y - BA = x - 65$

We know that $\overline{y} = B + \frac{\sum x_i}{n}$

(i.e) $\overline{x} = 65 + \frac{55}{10}$

(i.e) $\overline{x} = 70.5$.

and variance = σ^2

$$= \frac{1}{n}\sum d_i^2 - \left(\frac{\sum d_i}{n}\right)^2$$

$$= \frac{1}{10}(2521) - \left(\frac{55}{10}\right)^2$$

$$= 222$$

$\therefore$ standard deviation = σ_B = 14.9

Thus coefficient of variation for B = CV_B

$$= \frac{\sigma_B}{\bar{y}} \times 100$$

$$= \frac{14.9}{70.5} \times 100$$

$$= 21.135$$

Step 3 : Since $CV_A < CV_B$, the player A is more consistent.

Example 1. 24

A sample of five items is taken from the production of a firm. Length and weight of the five items are given below.

length (inches)	3	4	6	7	10
weight (ounces)	9	11	14	15	16

Calculate the coefficient of variation of these two characteristics and determine which of them is more variable.

Solution :

Step 1 : First we shall find coefficient of variation of length.

	x	$x-\bar{x}$	$(x-\bar{x})^2$
	3	-3	9
	4	-2	4
	6	0	0
	7	1	1
	10	4	16
Total	30		30

Space for Hint

Space for Hint

We know that $\bar{x} = \frac{\sum x_i}{n}$

(i.e) $\bar{x} = \frac{30}{5}$

(i.e) $\bar{x} = 6.$

and variance = σ^2

$$= \frac{1}{n}\sum(x_i - \bar{x})^2$$

$$= \frac{30}{5}$$

$$= 6$$

$\therefore$ standard deviation = $\sigma_A = 2.45$

Thus coefficient of variation for A

$$= \frac{\sigma_A}{\bar{x}} \times 100$$

$$= \frac{2.45}{6} \times 100$$

$$= 40.83$$

Step 2 : To find the coefficient of variation of weight.

	y	$y - \bar{y}$	$(y - \bar{y})^2$
	9	-4	16
	11	-2	4
	14	1	1
	15	2	4
	16	3	9
total	65		34

We know that $\bar{y} = \frac{\sum y_i}{n}$

(i.e) $\bar{x} = \frac{65}{5}$

(i.e) $\bar{x} = 13$

Space for Hint

and variance = σ^2

$$= \frac{1}{n}\sum(y_i - \bar{y})^2$$

$$= \frac{34}{5}$$

$$= 6.8$$

$\therefore$ Standard deviation = $\sigma_B = 2.61$

Thus coefficient of variation for B

$$= \frac{\sigma_B}{\bar{y}} \times 100$$

$$= \frac{2.61}{13} \times 100$$

$$= 20.08$$

Step 3 : Since coefficient of variation of length is greater than that of weight and hence length of the five items is more variable.

Check Your Progress

Find the standard deviation from the following frequency distribution.

Marks			No. of students
30	-	39	37
40	-	49	50
50	-	59	42
60	-	69	21
70	-	79	11
80	-	89	3

(Answer S.D. = 12.55)

Example 1. 25

Calculate the coefficient of variation for a series for which the following results are known $N = 50, \sum d_i = -10, \sum d_i^2 = 400$ where $d_i = x_i - 75$.

Space for Hint

Solution :

Given that $N = 50, \sum d_i = -10, \sum d_i^2 = 400$ and $d_i = x_i - 75$.

Here $A = 75$.

$$\therefore \ \bar{x} = A + \frac{\sum_{i=1}^{n} d_i}{N}$$

$$= 75 + \frac{(-10)}{50}$$

$$= 75 - 0.2$$

$$= 74.8$$

and standard deviation = σ

$$= \sqrt{\frac{1}{N}\sum_{i=1}^{n} d_i^2 - \left(\frac{1}{N}\sum_{i=1}^{n} d_i\right)^2}$$

$$= \sqrt{\frac{1}{50}(400) - \left(\frac{-10}{50}\right)^2}$$

$$= 2.82$$

Thus coefficient of variation = C.V.

$$= \frac{\sigma}{\bar{x}} \times 100$$

$$= \frac{2.82}{74.8} \times 100$$

$$= 3.77$$

Example 1. 26 :

Given $\sum x_i = 99, n = 9, \sum (x_i - 10)^2 = 79$. Find $\sum x_i^2$ and σ^2.

Solution : Given that $\sum x_i = 99, n = 9, \sum (x_i - 10)^2 = 79$.

Now $\sum (x_i - 10)^2 = 79$

$$\Rightarrow \sum x_i^2 - 20\sum x_i + \sum 100 = 79$$

$$\Rightarrow \sum x_i^2 - 20(99) + 9 \times 100 = 79$$

$$\Rightarrow \sum x_i^2 - 1980 + 900 = 79$$

$$\Rightarrow \sum x_i^2 = 1159.$$

Again $\sigma^2 = \frac{1}{n}\sum x_i^2 - \left(\frac{\sum x_i}{n}\right)^2$

$= \frac{1}{9}(1159) - \left(\frac{99}{9}\right)^2$

$= \frac{70}{9}$

Thus $\sum x_i^2 = 1159$ and $\sigma^2 == \frac{70}{9}$

Theorem 1. 4 (Variance of a combined set)

Let n_1, n_2 be the number of items in two groups having means $\overline{x}_1, \overline{x}_2$ and variances σ_1^2, σ_2^2 respectively. Then $\overline{x} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2}{n_1 + n_2}$ is the mean and

$\sigma^2 = \frac{1}{n_1 + n_2}\left[n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)\right]$ is the variance of the combined group where $d_1 = \overline{x}_1 - \overline{x}$, $d_2 = \overline{x}_2 - \overline{x}$.

Proof :

Given that $n_1, \overline{x}_1, \sigma_1^2$ and $n_2, \overline{x}_2, \sigma_2^2$ are the number of items, mean and variances of two groups.

Step 1 : To find the mean of the combined group.

Now the sum of the observations of the first group = $n_1\ \overline{x}_1$

and the sum of the observations of the first group = $n_2\ \overline{x}_2$.

$\therefore$ the sum of the observations in the combined group = $n_1\ \overline{x}_1 + n_2\ \overline{x}_2$.

and the total number of observation in the combined group = $n_1 + n_2$.

Thus the mean of the combined group = $\overline{x}$

$$= \frac{n_1\overline{x}_1 + n_2\overline{x}_2}{n_1 + n_2}$$

This completes **step 1.**

Step 2 : to find the variance of combined group.

Now variance of combined group = σ^2

$$= \frac{1}{n_1 + n_2}\sum_{i=1}^{n_1+n_2}(x_i - \overline{x})^2$$

Space for Hint

Space for Hint

$$= \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1}(x_i - \bar{x})^2 + \sum_{i=n_1+1}^{n_1+n_2}(x_i - \bar{x})^2\right] \text{------------------ (1.6)}$$

Now $\sum_{i=1}^{n_1}(x_i - \bar{x})^2 = \sum_{i=1}^{n_1}(x_i - \bar{x}_1 + \bar{x}_1 - \bar{x})^2$

$$= \sum_{i=1}^{n_1}\left[(x_i - \bar{x}_1)^2 + (\bar{x}_1 - \bar{x})^2 + 2(\bar{x}_1 - \bar{x})(x_i - \bar{x})\right]$$

$$= \sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 + \sum_{i=1}^{n_1} d_1^2 + 0 \text{ where } d_1 = \bar{x}_1 - \bar{x}$$

$$= n_1\sigma_1^2 + n_1 d_1^2$$

$$= n_1(\sigma_1^2 + d_1^2) \text{ --- (1.7)}$$

Similarly $\sum_{n_1+1}^{n_1+n_2}(x_i - \bar{x})^2 = n_2(\sigma_2^2 + d_2^2)$ where $d_2 = \bar{x}_2 - \bar{x}$ ------ (1.8)

From (1.6), (1.7) and (1.8) we have,

$$\sigma^2 = \frac{1}{n_1 + n_2}\left[n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)\right].$$

This proves the theorem.

Note : Variance of combined group can also be obtained from

$$\sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2} + \frac{n_1 \cdot n_2}{(n_1 + n_2)^2}(\bar{x}_1 - \bar{x}_2)^2$$

Proof of the note :

We know that the mean and variance of the combined group are

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} \text{ and}$$

$$\sigma^2 = \frac{1}{n_1 + n_2}\left[n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)\right] \text{ ---------------------- (1.9)}$$

where $d_1 = \bar{x}_1 - \bar{x}$, $d_2 = \bar{x}_2 - \bar{x}$.

Now $d_1 = \bar{x}_1 - \bar{x}$

$$= \bar{x}_1 - \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

$$= \frac{\cancel{n_1\bar{x}_1} + n_2\bar{x}_1 - \cancel{n_1\bar{x}_1} - n_2\bar{x}_2}{n_1 + n_2}$$

$$= \frac{n_2(\bar{x}_1 - \bar{x}_2)}{n_1 + n_2}$$

$$\therefore\ d_1^2 = \left(\frac{n_2}{n_1 + n_2}\right)^2 (\bar{x}_1 - \bar{x}_2)^2$$

Similarly $d_2^2 = \left(\frac{n_1}{n_1 + n_2}\right)^2 (\bar{x}_1 - \bar{x}_2)^2$

Now (1.9) $\Rightarrow \sigma^2 = \frac{1}{n_1 + n_2}\left[n_1\sigma_1^2 + n_1 d_1^2 + n_2\sigma_2^2 + n_2 d_2^2\right]$

$$= \frac{1}{n_1 + n_2}\left[n_1\sigma_1^2 + n_1\left(\frac{n_2}{n_1 + n_2}\right)^2 (\bar{x}_1 - \bar{x}_2)^2 \right.$$

$$\left. + n_2\sigma_2^2 + n_2\left(\frac{n_1}{n_1 + n_2}\right)^2 (\bar{x}_1 - \bar{x}_2)^2\right]$$

$$= \frac{1}{n_1 + n_2}\left[n_1\sigma_1^2 + n_2\sigma_2^2 + \frac{n_1 . n_2}{(n_1 + n_2)^2}(\bar{x}_1 - \bar{x}_2)^2\right]$$

$$= \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2} + \frac{n_1 \cdot n_2}{(n_1 + n_2)^2}(\bar{x}_1 - \bar{x}_2)^2$$

(i.e) $\sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2} + \frac{n_1 \cdot n_2}{(n_1 + n_2)^2}(\bar{x}_1 - \bar{x}_2)^2$

This proves the note.

Example 1. 27 :

The means of the two samples of sizes 50 and 100 respectively are 54.1 and 50.3 and the standard deviations are 8 and 7. Obtain the mean and standard deviation of the sample of size 150 obtained by combining the two samples.

Solution :

Given that $n_1 = 50,\ n_2 = 100,$

$\bar{x}_1 = 54.1,\ \bar{x}_2 = 50.3,$

$\sigma_1 = 8,\ \sigma_2 = 7.$

Step 1 : To find the mean of the combined group.

We know that the combined mean is $\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$

Space for Hint

Space for Hint

(i.e) $\bar{x} = \dfrac{50 \times 54.1 + 100 \times 50.3}{50 + 100}$

(i.e) $\bar{x} = 51.57$

Step 2 : To find the standard deviation of the combined group.

We know that $\sigma^2 = \dfrac{n_1 \sigma_1^2 + n_2 \sigma_2^2}{n_1 + n_2} + \dfrac{n_1 \cdot n_2}{(n_1 + n_2)^2}(\bar{x}_1 - \bar{x}_2)^2$

$= \dfrac{50 \times 64 + 100 \times 49}{50 + 100} + \dfrac{50 \times 100}{(50 + 100)^2}(54.1 - 50.3)^2$

$= 54 + \dfrac{2}{9} \times 14.44$

$= 57.21$

Thus $\sigma = 7.564$

Hence the standard deviation of the combined group is 7.565

Example 1. 28 :

The first group of the sample has 100 items with mean 15 and standard deviation 3. if the whole group has 250 items with mean 15.6 and standard deviation $\sqrt{13.44}$ find the standard deviation of the second group.

Solution :

Given that $n_1 = 100, n_2 = ?, n = 250,$

$\bar{x}_1 = 15, \bar{x}_2 = ?, \bar{x} = 15.6,$

$\sigma_1 = 3, \sigma_2 = ?, \sigma = \sqrt{13.44}.$

Now $n = 250$

(i.e) $n_1 + n_2 = 250$

(i.e) $100 + n_2 = 250$

$\therefore\ n_2 = 150$

Now $\bar{x} = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$

(i.e) $15.6 = \dfrac{100 \times 15 + 150\bar{x}_2}{250}$

(i.e) $150\bar{x}_2 = 15.6 \times 250 - 1500$

(i.e) $150\bar{x}_2 = 2400$

Space for Hint

(i.e) $\overline{x}_2 = 16$

(i.e) mean of the second group is 16.

Again we know that $\sigma^2 = \dfrac{n_1 \sigma_1^2 + n_2 \sigma_2^2}{n_1 + n_2} + \dfrac{n_1 \cdot n_2}{(n_1 + n_2)^2}(\overline{x}_1 - \overline{x}_2)^2$.

$$\text{(i.e)}\ 13.44 = \frac{100 \times 9 + 150\sigma_2}{250} + \frac{100 \times 150}{(100 \times 150)^2}(15 - 16)^2$$

$$\text{(i.e)}\ 13.44 = \frac{900 + 150\sigma_2^2}{250} + 0.24$$

(i.e) $150\sigma_2^2 = 13.2 \times 250 - 900$

(i.e) $\sigma_2^2 = 16$

$\therefore\ \sigma_2 = 4$

(i.e) standard deviation of the second group is 4.

Example 1. 29

Find the missing information from the following data

	Group I	Group II	Group III	Combined group
Number	50	?	90	200
S.D.	6	7	?	7.745
Mean	113	?	115	116

Solution :

Given that $n_1 = 50,\ n_2 = ?,\ n_3 = 90,\ n = 200,$

$\overline{x}_1 = 113,\ \overline{x}_2 = ?,\ \overline{x}_3 = 115,\ \overline{x} = 116,$

$\sigma_1 = 6,\ \sigma_2 = 7,\ \sigma_3 = ?,\ \sigma = 7.745.$

Step 1 : To find n_2

Here $n = 200$

(i.e) $n_1 + n_2 + n_3 = 200$

(i.e) $50 + n_2 + 90 = 200$

(i.e) $n_2 = 60$

This completes step 1.

Space for Hint

Step 2 : To find $\overline{x}_2$

We know that $\overline{x} = \dfrac{n_1\overline{x}_1 + n_2\overline{x}_2 + n_3\overline{x}_3}{n_1 + n_2 + n_3}$

(i.e) $116 = \dfrac{50 \times 113 + 60 \times \overline{x}_2 + 90 \times 115}{200}$

(i.e) $60\overline{x}_2 = 7200$

$\therefore\ \overline{x}_2 = 120$

This completes step 2.

Step 3 : To find σ_3.

We know that

$$\sigma^2 = \frac{1}{n_1 + n_2 + n_3}\left[n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2) + n_3(\sigma_3^2 + d_3^2)\right]$$

Now $d_1 = \overline{x}_1 - \overline{x}$

$= 116 - 113$

$= 3,$

$d_2 = \overline{x}_2 - \overline{x}$

$= 116 - 120$

$= -4$

and $d_3 = \overline{x}_3 - \overline{x}$

$= 116 - 115$

$= 1$

Now $\sigma^2 = \dfrac{1}{n_1 + n_2 + n_3}\left[n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2) + n_3(\sigma_3^2 + d_3^2)\right]$

(i.e) $(7.745)^2 = \dfrac{1}{200}\left[50 \times (36 + 9) + 60 \times (49 + 16) + 90 \times (\sigma_3^2 + 1)\right]$

(i.e) $11997.005 = 2250 + 3900 + 90 \times (\sigma_3^2 + 1)$

(i.e) $90 \times (\sigma_3^2 + 1) = 5847.005$

(i.e) $\sigma_3^2 = 63.97$

Thus $\sigma = 7.998$

Hence the standard deviation of the combined group is 7.998

Check Your Progress

(1) The scores of two cricketers a and B in 10 innings are given below. Find who is a better run getter and who is more consistent player?

A's score	40	25	19	80	38	8	67	121	66	76
B's score	28	70	31	0	14	111	66	31	25	4

(Answer : A is better run getter and A is the consistent player)

(2) The mean and standard deviation of 200 items are found to be 60 and 20. If at the time of calculation two items are wrongly taken as 3 and 67 instead of 13 and 17, finds the correct mean and standard deviation.

(Answer correct mean and standard deviation are 59.8 and 20.09 respectively)

Merits :

(1) It is rigidly defined

(2) It is based on all the observations of the series.

(3) Least affected by fluctuations of sampling.

Demerits :

(1) It cannot be comparing the dispersion of two or more series of observations given in different units.

(2) Difficult to understand and compute

(3) Gives more weightage to extreme values.

1.3 Moments

Definition : The r^{th} moment about any value A of a distribution is defined as $\frac{\Sigma f_i (x_i - A)^r}{N}$ and it is denoted as μ_r'.

$$\text{(i.e)}\ \mu_r' = \frac{\Sigma f_i (x_i - A)^r}{N}$$

Definition : The r^{th} moment about origin of a distribution is defined as

$\mu_r' = \frac{\Sigma f_i x_i^r}{N}$.

$$\text{(i.e)}\ \mu_r' = \frac{\Sigma f_i x_i^r}{N}.$$

Space for Hint

Space for Hint

Definition : The r^{th} moment about mean of a distribution it is denoted as μ_r and is defined as $\frac{\sum f_i (x_i - \bar{x})^r}{N}$. It is also called as r^{th} central moment.

$$\text{(i.e)}\ \mu_r = \frac{\sum f_i (x_i - \bar{x})^r}{N}$$

Note :

(1) The first moment about origin coincides with the arithmetic mean of the frequency distribution.

(2) $\mu_1 = \frac{\sum f_i (x_i - \bar{x})^1}{N} = 0$

(3) $\mu_2 = \frac{\sum f_i (x_i - \bar{x})^2}{N}$ which is variance of the distribution.

(4) $\mu_1' = \frac{1}{N} \sum f_i (x_i - A)$

$$= \frac{1}{N}(\sum f_i x_i) - \frac{1}{N} A \sum f_i$$

$$= \bar{x} - A$$

(i.e) $\mu_1' = \bar{x} - A$

$$\therefore \bar{x} = A + \mu_1'$$

Example 1. 30 :

Prove that

$$\mu_r = \mu_r' - {}^rC_1 \mu_{r-1}' \mu_1' + {}^rC_2 \mu_{r-2}' (\mu_1')^2 - \cdots + (-1)^{r-1}(r-1)(\mu_r')^r$$

Proof :

L.H.S. = μ_r

$= \frac{1}{N} \sum f_i (x_i - \bar{x})^r$

$= \frac{1}{N} \sum f_i (x_i - A + A - \bar{x})^r$

$= \frac{1}{N} \sum f_i (x_i - A - \mu_1')^r$

$$=\frac{1}{N}\Big[\Sigma f_i(x_i - A)^r - {}^rC_1\,\mu_1'\,\Sigma f_i(x_i - A)^{r-1} + {}^rC_2\,(\mu_1')^2\,\Sigma f_i(x_i - A)^{r-2}$$

$$+\cdots + {}^rC_{r-1}(-\mu_1')^{r-1}\,\Sigma f_i(x_i - A) + {}^rC_r(-\mu_1')^r\,\Sigma f_i\Big]$$

$$= \mu_r' - {}^rC_1\,\mu_{r-1}'\,\mu_1' + {}^rC_2\,\mu_{r-2}'\,(\mu_1')^2 - \cdots + (-1)^{r-1}(r-1)(\mu_r')^r$$

Thus $\mu_r = \mu_r' - {}^rC_1\,\mu_{r-1}'\,\mu_1' + {}^rC_2\,\mu_{r-2}'\,(\mu_1')^2 - \cdots + (-1)^{r-1}(r-1)(\mu_r')^r$

This proves the problem.

Note : putting $n = 2,3,4$ in the above problem we get,

(i) $\mu_2 = \mu_2' - (\mu_1')^2$

(ii) $\mu_3 = \mu_3' - 3\mu_2'\,\mu_1' + 2(\mu_1')^3$

(iii) $\mu_4 = \mu_4' - 4\mu_3'\,\mu_1' + 6\mu_2'\,(\mu_1')^2 - 3(\mu_1')^4$

Example 1. 31 :

Prove that $\mu_r' = \mu_r + {}^rC_1\,\mu_{r-1}\,\mu_1' + {}^rC_2\,\mu_{r-2}\,(\mu_1')^2 - \cdots + (\mu_r)^r$

Proof :

L.H.S. $= \mu_r'$

$$= \frac{1}{N}\Sigma f_i(x_i - A)^r$$

$$= \frac{1}{N}\Sigma f_i(x_i - \bar{x} + \bar{x} - A)^r$$

$$= \frac{1}{N}\Sigma f_i(x_i - \bar{x} - \mu_1')^r$$

$$=\frac{1}{N}\Big[\Sigma f_i(x_i - \bar{x})^r - {}^rC_1\,\mu_1'\,\Sigma f_i(x_i - \bar{x})^{r-1} + {}^rC_2\,(\mu_1')^2\,\Sigma f_i(x_i - \bar{x})^{r-2}$$

$$+\cdots + {}^rC_r(-\mu_1')^r\,\Sigma f_i\Big]$$

$$= \mu_r + {}^rC_1\,\mu_{r-1}\,\mu_1' + {}^rC_2\,\mu_{r-2}\,(\mu_1')^2 - \cdots + (\mu_r)^r$$

Thus $\mu_r' = \mu_r + {}^rC_1\,\mu_{r-1}\,\mu_1' + {}^rC_2\,\mu_{r-2}\,(\mu_1')^2 - \cdots + (\mu_r)^r$

This proves the problem.

Note : putting $n = 2,3,4$ in the above problem we get,

(i) $\mu_2' = \mu_2 + (\mu_1')^2$

(ii) $\mu_3' = \mu_3 + 3\mu_2\,\mu_1' + (\mu_1')^3$

(iii) $\mu_4' = \mu_4 + 4\mu_3\,\mu_1' + 6\mu_2\,(\mu_1')^2 + (\mu_1')^4$

Space for Hint

Space for Hint

Important note :

If $u_i = \dfrac{x_i - A}{c}$, then the r^{th} moment of the variable x_i is

$$\mu_r = c^r\left[\frac{1}{N}\Sigma f_i (u_i - \bar{u})^r\right].$$

(i.e) μ_r with respect to $x_i = c^r \times \mu_r$ with respect to u_i.

Definition : Karl Pearson's β and γ are defined as

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2}.$$

$$\gamma_1 = \sqrt{\beta_1} \text{ and } \gamma_2 = \beta_2 - 3.$$

Example 1. 32

Find the first four moments to the following data

Wages in Rs.	:	10	11	12	13	14	15
Frequency	:	2	4	10	8	5	1

Solution :

Given that

Wages in Rs.	:	10	11	12	13	14	15
Frequency	:	2	4	10	8	5	1

First we shall form the following table to find $N, \Sigma fx, \Sigma fx^2, \Sigma fx^3, \Sigma fx^4$.

Wages in Rs. (x)	Frequency (f)	fx	fx^2	fx^3	fx^4
10	2	20	200	2000	20000
11	4	44	484	5324	58564
12	10	120	1440	17280	207360
13	8	104	1352	17576	228488
14	5	70	980	13720	192080
15	1	15	225	3375	50625
total	30	373	4681	59275	757117

Now $\mu_1' = \dfrac{1}{N}\Sigma fx$

Space for Hint

$$= \frac{373}{30}$$

$$= 12.43$$

and $\mu_2' = \frac{1}{N}\sum fx^2$

$$= \frac{4681}{30}$$

$$= 156.03$$

and $\mu_3' = \frac{1}{N}\sum fx^3$

$$= \frac{59275}{30}$$

$$= 1975.83$$

and $\mu_4' = \frac{1}{N}\sum fx^4$

$$= \frac{757117}{30}$$

$$= 25237.23$$

Now $\mu_1 = 0$

and $\mu_2 = \mu_2' - (\mu_1')^2$

$$= 156.03 - (12.43)^2$$

$$= 1.4456$$

and $\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$

$$= 1975.83 - 3 \times 156.03 \times 12.43 + 2 \times (12.43)^3$$

$$= -0.1273$$

and $\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$

$$= 25237.23 - 4(1975.83)(12.43) + 6(156.03)(12.43)^2 - 3(12.43)^3$$

$$= 5.39$$

Thus $\mu_1 = 0$, $\mu_2 = 1.4456$, $\mu_3 = -0.1273$ and $\mu_4 = 5.39$

Space for Hint

Example 1. 33 :

Find the first four moments to the following data

Size	:	6	7	8	9	10	11	12
Frequency	:	3	6	9	13	8	5	4

Solution :

Given that

Size	:	6	7	8	9	10	11	12
Frequency	:	3	6	9	13	8	5	4

Now

Size	Frequency (f)	fx	$f(x-\overline{x})^2$	$f(x-\overline{x})^3$	$f(x-\overline{x})^4$
6	3	18	27	-81	243
7	6	42	24	-48	96
8	9	72	9	-9	9
9	13	117	0	0	0
10	8	80	8	8	8
11	5	55	20	40	80
12	4	48	36	108	324
total	48	432	124	18	760

Now $\overline{x} = \frac{1}{N}\Sigma fx$

$= \frac{432}{48}$

$= 9$

Now $\mu_1 = 0$

and $\mu_2 = \frac{1}{N}\Sigma f(x-\overline{x})^2$

$= \frac{124}{48}$

$= 2.583$

and $\mu_3 = \frac{1}{N}\Sigma f(x-\bar{x})^3$

$= \frac{18}{48}$

$= 0.375$

and $\mu_4 = \frac{1}{N}\Sigma f(x-\bar{x})^4$

$= \frac{760}{48}$

$= 15.83$

Thus $\mu_1 = 0$, $\mu_2 = 2.583$, $\mu_3 = 0.375$ and $\mu_4 = 15.83$.

Example 1. 34

Calculate the first four moments of a distribution. Also find the values of β_1 and β_2.

x	:	0	1	2	3	4	5	6	7	8
f	:	5	10	15	20	25	20	15	10	5

Solution :

Given that

x	:	0	1	2	3	4	5	6	7	8
f	:	5	10	15	20	25	20	15	10	5

Now

x	f	fx	$f(x-\bar{x})^2$	$f(x-\bar{x})^3$	$f(x-\bar{x})^4$
0	5	0	80	-320	1280
1	10	10	90	-270	810
2	15	30	60	-120	240
3	20	60	20	-20	20
4	25	100	0	0	0
5	20	100	20	20	20
6	15	90	60	120	240
7	10	70	90	270	810
8	5	40	80	320	1280
total	125	500	500	0	4700

Space for Hint

Space for Hint

Now $\bar{x} = \frac{1}{N}\Sigma fx$

$= \frac{500}{125}$

$= 4$

Now $\mu_1 = 0$

and $\mu_2 = \frac{1}{N}\Sigma f(x-\bar{x})^2$

$= \frac{500}{125}$

$= 4$

and $\mu_3 = \frac{1}{N}\Sigma f(x-\bar{x})^3$

$= \frac{0}{125}$

$= 0$

and $\mu_4 = \frac{1}{N}\Sigma f(x-\bar{x})^4$

$= \frac{4700}{125}$

$= 37.6$

Thus $\mu_1 = 0$, $\mu_2 = 4$, $\mu_3 = 0$ and $\mu_4 = 37.6$.

Again $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$

$= \frac{0}{4^3}$

$= 0$

and $\beta_2 = \frac{\mu_4}{\mu_2^2}$

$= \frac{37.6}{4^2}$

$= 2.35$

Example 1. 35 :

Calculate the first four moments of a distribution about $x = 4$ and hence find the moments about the mean of the distribution. Also find the values of β_1 and β_2.

x	:	0	1	2	3	4	5	6	7	8	9	10
f	:	5	10	30	70	140	200	140	70	30	10	5

Solution :

Given that

x	:	0	1	2	3	4	5	6	7	8	9	10
f	:	5	10	30	70	140	200	140	70	30	10	5

Now we shall find the values of $\frac{1}{N}\Sigma f(x-\bar{x})$, $\frac{1}{N}\Sigma f(x-\bar{x})^2$, $\frac{1}{N}\Sigma f(x-\bar{x})^3$ and $\frac{1}{N}\Sigma f(x-\bar{x})^4$ using the following table.

x	f	$f(x-4)$	$f(x-4)^2$	$f(x-4)^3$	$f(x-4)^4$
0	5	-20	80	-320	1280
1	10	-30	90	-270	810
2	30	-60	120	-240	480
3	70	-70	70	-70	70
4	140	0	0	0	0
5	200	200	200	200	200
6	140	280	560	1120	2240
7	70	210	630	1890	5670
8	30	120	480	1920	7680
9	10	50	250	1250	6250
10	5	30	180	1080	6480
total	710	710	2660	6560	31160

Space for Hint

Space for Hint

Now $\mu_1' = \frac{1}{N}\Sigma f(x-4)$

$= \frac{710}{710}$

$= 1$

and $\mu_2' = \frac{1}{N}\Sigma f(x-4)^2$

$= \frac{2660}{710}$

$= 3.746$

and $\mu_3' = \frac{1}{N}\Sigma f(x-4)^3$

$= \frac{6560}{710}$

$= 9.239$

and $\mu_4' = \frac{1}{N}\Sigma f(x-4)^4$

$= \frac{31160}{710}$

$= 43.887$

Now $\mu_1 = 0$

and $\mu_2 = \mu_2' - (\mu_1')^2$

$= 3.746 - (1)^2$

$= 2.746$

and $\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$

$= 9.239 - 3(3.746)(1) + 2(1)^3$

$= 0$

and $\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$

$= 43.887 - 4(9.239)(1) + 6(3.746)(1)^2 - 3(1)^4$

$= 26.407$

Thus $\mu_1 = 0$, $\mu_2 = 2.746$, $\mu_3 = 0$ and $\mu_4 = 26.407$

Space for Hint

Again $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$

$= \frac{0}{(2.746)^3}$

$= 0$

and $\beta_2 = \frac{\mu_4}{\mu_2^2}$

$= \frac{26.407}{(2.746)^2}$

$= 3.502$

Example 1. 36 :

The first four moments of a distribution about $x = 5$ are 2, 20, 40 and 50. Show that the mean =7, variance = 16, $\mu_3 = -64$, $\mu_4 = 162$, $\beta_1 = 1$ and $\beta_2 = 0.63$.

Solution :

Given that $\mu_1' = 2$, $\mu_2' = 20$, $\mu_3' = 40$, $\mu_4' = 50$ and $A = 5$

Now mean $= \bar{x} = A + \mu_1'$

$= 5 + 2$

$= 7,$

and $\mu_2 = \mu_2' - (\mu_1')^2$

$= 20 - (2)^2$

$= 16$

and $\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$

$= 40 - 3(20)(2) + 2(2)^3$

$= -64$

and $\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$

$= 50 - 4(40)(2) + 6(20)(2)^2 - 3(2)^4$

$= 162$

Thus $\mu_2 = 16$, $\mu_3 = -64$ and $\mu_4 = 162$

Space for Hint

Again $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$

$= \frac{(-64)^2}{(16)^3}$

$= 1$

and $\beta_2 = \frac{\mu_4}{\mu_2^2}$

$= \frac{162}{(16)^2}$

$= 0.63$

This proves the problem.

Check Your Progress

(1) The first four moments of a distribution about $x = 4$ are $-1.5, 17, -30, 108$. Find the first four moments about (i) mean, (ii) the origin, (iii) $x = 2$ and (iv) β_1 *and* β_2.

(2) For a distribution the mean is 10, variance is 16, γ_1 is 1 and β_2 is 4. Find the first four moments about the origin.

(3) First four moments of a distribution about $x = 2$ are 1, 2.5, 5.5 and 16. Calculate the four moments about the (i) mean and (ii) zero.

(Answer : moments about mean are 0, 1.5, 0, 6

moments about origin are 3,10.5, 40.5, 168

1. 4 Skewness and Kurtosis

If the values of a variable are distributed symmetrically about its mean then $\beta_1 = 0$.

If a distribution is asymmetric then we say that the distribution is skewed distribution. Hence skewness means lack of symmetry.

If $\beta_1 > 0$ then we say the distribution is positively skewed distribution , otherwise it is called negative skewness.

Note that for a symmetrical distribution mean = median = mode.

Karl Pearson's coefficient of skewness

Karl Pearson's coefficient of skewness is defined as $\frac{mean - mode}{S.D}$ or $\frac{3(mean - median)}{S.D}$.

Bowley's coefficient of skewness

Bowley's coefficient of skewness is defined as $\frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$.

Kurtosis :

Kurtosis is the degree of peakness of a distribution. Kurtosis gives the idea bout the flatness or peakness of a frequency curve.

For a normal curve $\beta_2 = 3$ or $\gamma_1 = 0$ called mesodutic.

For a curve $\beta_2 < 3$ or $\gamma_1 < 0$ called platykurtic.

For a curve $\beta_2 > 3$ or $\gamma_1 > 0$ called leptokurtic.

Example 1. 37

Find the Karl Pearson's coefficient of skewness for the following distribution

age	students	age	students
10 - 12	4	18 - 20	20
12 - 14	10	20 - 22	14
14 - 16	16	22 - 24	6
16 - 18	30	Total	100

Solution :

Step 1 : To find mean

Age	No. of students	midvalue	$d = \frac{x-17}{2}$	fd	fd^2	cf
10 - 12	4	11	-3	-12	36	4
12 - 14	10	13	-2	-20	40	14
14 - 16	16	15	-1	-16	16	30
16 - 18	30	17	0	0	0	60
18 - 20	20	19	1	20	20	80
20 - 22	14	21	2	28	56	94
22 - 24	6	23	3	18	54	100
Total	100			18	222	

Space for Hint

Take $A = 17$

Here $c = 2$

Thus mean = $\bar{x}$

$$= A + \frac{\sum fd}{\sum f} \times c$$

$$= 17 + \frac{18}{100} \times 2$$

$$= 17.36$$

Step 2 : To find the median

We know that median = $l + \frac{N/2 - f_1}{f_2} \times c$

$$= 16 + \frac{50 - 30}{30} \times 2$$

$$= 16 + \frac{20}{30} \times 2$$

$$= 17.33$$

Step 3 : To find standard deviation

We know that S.D. = σ

$$= \sqrt{\frac{1}{N} \times fd^2 - \left(\frac{\sum fd}{N}\right)^2} \times c$$

$$= \sqrt{\frac{222}{100} - \left(\frac{18}{100}\right)^2} \times 2$$

$$= 2.96$$

Step 4 : To find the Karl Pearson's coefficient of skewness.

Now Karl Pearson's coefficient of skewness

$$= \frac{3(mean - median)}{S.D}$$

$$= \frac{3(17.36 - 17.33)}{2.96}$$

$$= 0.03$$

Example 1. 38 :

Find the Karl Pearson's coefficient of skewness for the following distribution

Marks	Students	Marks	Students
0 - 10	10	40 - 50	10
10 - 20	40	50 - 60	40
20 - 30	20	60 - 70	16
30 - 40	0	70 80	14

Space for Hint

Solution :

Step 1 : To find mean

Marks	Students	Midvalue	$d = \dfrac{x-45}{10}$	fd	fd^2	cf
0 -10	10	5	-4	-40	160	10
10-20	40	15	-3	-120	360	50
20-30	20	25	-2	-40	80	70
30-40	0	35	-1	0	0	70
40 50	10	45	0	0	0	80
50-60	40	55	1	40	40	120
60-70	16	65	2	32	64	136
70-80	14	75	3	42	126	150
Total	150			-86	830	

Take $A = 45$

Here $c = 10$

Thus mean = $\overline{x}$

$$= A + \frac{\sum fd}{\sum f} \times c$$

$$= 45 + \frac{-86}{150} \times 10$$

$$= 39.27$$

Space for Hint

Step 2 : To find the median

We know that median $= l + \dfrac{N/2 - f_1}{f_2} \times c$

$$= 40 + \frac{75-70}{10} \times 10$$

$$= 40 + 5$$

$$= 45$$

Step 3 : To find standard deviation

We know that S.D. $= \sigma$

$$= \sqrt{\frac{1}{N} \times fd^2 - \left(\frac{\Sigma fd}{N}\right)^2} \times c$$

$$= \sqrt{\frac{830}{150} - \left(\frac{-86}{150}\right)^2} \times 10$$

$$= 22.81$$

Step 4 : To find the Karl Pearson's coefficient of skewness.

Now Karl Pearson's coefficient of skewness

$$= \frac{3(mean - median)}{S.D}$$

$$= \frac{3(39.27-45)}{22.81}$$

$$= -0.76$$

Example 1. 39

Find the Bowley's coefficient of skewness for the following distribution

Marks	Students	Marks	Students
1 - 5	20	21 - 25	48
6 - 10	27	26 - 30	53
11 - 15	29	31 - 35	70
16 - 20	38	Total	285

Solution :

Step 1 : To find mean

Marks			Frequency	cf
0.5	-	5.5	20	20
5.5	-	10.5	27	47
10.5	-	15.5	29	76
15.5	-	20.5	38	114
20.5	-	25.5	48	162
25.5	-	30.5	53	215
30.5	-	35.5	70	285

Step 2 : To find the Q_1

We know that $Q_1 = l + \frac{N/4 - f_1}{f_2} \times c$

Here $\frac{N}{4} = \frac{285}{4} = 71.25$

Thus $Q_1 = 10.5 + \frac{71.25 - 47}{29} \times 5$

(i.e) $Q_1 = 14.68$

Step 3 : To find the median

We know that median $= l + \frac{N/2 - f_1}{f_2} \times c$

Here $\frac{N}{2} = \frac{285}{2} = 142.5$

Thus median $= Q_2 = 20.5 + \frac{142.5 - 114}{48} \times 5$

(i.e) $Q_2 = 23.47$

Step 4 : To find Q_3

We know that $Q_3 = l + \frac{\frac{3N}{4} - f_1}{f_2} \times c$

Here $\frac{3N}{4} = \frac{3 \times 285}{4} = 213.75$

Space for Hint

Space for Hint

Thus $Q_3 = 25.5 + \frac{213.75 - 162}{53} \times 5$

(i.e) $Q_3 = 30.38$

Step 5 : To find the Bowley's coefficient of skewness.

Now Bowley's coefficient of skewness

$$= \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

$$= \frac{30.38 + 14.68 - 2 \times 23.47}{30.38 - 14.68}$$

$$= -0.12$$

Example 1. 40 :

Find the Bowley's coefficient of skewness for the following distribution

Annual sales	No. of firms
0 - 100	20
100 - 250	50
250 - 500	69
500 - 1250	30
1250 - 2500	25
2500 - 5000	19

Solution :

Step 1 : To find mean

Annual sales	No. of firms	cf
0 - 100	20	20
100 - 250	50	70
250 - 500	69	139
500 - 1250	30	169
1250 2500	25	194
2500 5000	19	213

Step 2 : To find the Q_1

We know that $Q_1 = l + \frac{N/4 - f_1}{f_2} \times c$

Here $\frac{N}{4} = \frac{213}{4} = 53.25$

Thus $Q_1 = 100 + \frac{53.25 - 20}{50} \times 150$

(i.e) $Q_1 = 199.75$

Step 3 : To find the median

We know that median $= l + \frac{N/2 - f_1}{f_2} \times c$

Here $\frac{N}{2} = \frac{213}{2} = 106.5$

Thus median $= Q_2 = 250 + \frac{106.5 - 70}{69} \times 250$

(i.e) $Q_2 = 382.25$

Step 4 : To find Q_3

We know that $Q_3 = l + \frac{\frac{3N}{4} - f_1}{f_2} \times c$

Here $\frac{3N}{4} = \frac{3 \times 213}{4} = 159.75$

Thus $Q_3 = 500 + \frac{159.75 - 139}{30} \times 750$

(i.e) $Q_3 = 1018.75$

Step 5 : To find the Bowley's coefficient of skewness.

Now Bowley's coefficient of skewness

$$= \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

$$= \frac{1018.75 + 199.75 - 2 \times 382.25}{1018.75 - 199.75}$$

$$= 0.554$$

Space for Hint

Space for Hint

Check Your Progress :

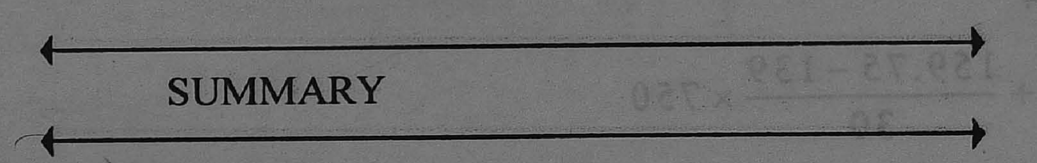
(1) Find the coefficient of variation of a frequency distribution given that its mean is 120, mode is 123 and Karl Pearson's coefficient of skewness is -0.3.

(2) Calculate the Karl Pearson's coefficient of skewness from the following distribution :

Marks	No. of students
more than 0	150
more than 10	140
more than 20	100
more than 30	80
more than 40	80
more than 50	70
more than 60	30
more than 70	14
more than 80	0

(3) The coefficient of skewness for a certain distribution based on the quartiles is -0.8. If the sum of the upper and lower quartiles is 100.7 and median is 55.35, find the distribution on the basis of the upper and lower quartiles.

(Answer : $Q_1 = 8$, $Q_3 = 20$)



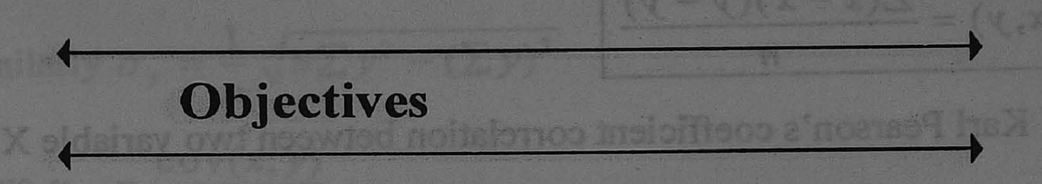
SUMMARY

In this unit we have learned the method of finding mean, median, mode, geometric mean, harmonic mean, range, mean deviation, quartile deviation, standard deviation, skewness and kurtosis.

Unit II

Space for Hint

CORRELATION, REGRESSION



Objectives

In this unit, we are going to discuss how to find the correlation between two variables and then to find the relationship between two variables.

After the completion of this unit one may able to fit

- Correlation coefficient between two variables.
- Regression equations.
- Rank correlation coefficient.

Introduction

Correlation is a statistical measure for finding the degree of association between two or more variables. Here "association" mean that the tendency of the variables moves together. If two variables x and y are so related that movements in one, tend to be accomplished by the corresponding movement in the other variable, then we say that two variables are correlated.

2. 1 Correlation coefficient

Let x be a variable having the values $x_1, x_2, x_3, \cdots, x_n$ and y be the second variable having the values $y_1, y_2, y_3, \cdots, y_n$. If there is a change in one variable corresponding to a change in the other variable we say that the variables are correlated.

If the two variables deviate in the same in the same direction the correlation is said to be direct or positive. If they always deviate in the opposite direction the correlation is said to be inverse or negative.

Space for Hint

Definition : The covariance between two variable X and Y is defined by $\frac{\sum(x-\bar{x})(y-\bar{y})}{n}$.

(i.e) $$\boxed{\operatorname{cov}(x,y)=\frac{\sum(x-\bar{x})(y-\bar{y})}{n}}$$

Definition : Karl Pearson's coefficient correlation between two variable X and Y is denoted by γ_{xy} and is defined by $\frac{\operatorname{cov}(x,y)}{\sigma_x\sigma_y}$.

(i.e) $$\gamma_{xy}=\frac{\operatorname{cov}(x,y)}{\sigma_x\sigma_y}$$

(i.e) $$\boxed{\gamma_{xy}=\frac{\sum(x-\bar{x})(y-\bar{y})}{n\sigma_x\sigma_y}}$$

Note : Two variables X and Y are independent if $\gamma_{xy}=0$

Theorem 2. 1

$$\gamma_{xy}=\frac{n\sum xy-(\sum x)(\sum y)}{\sqrt{n\sum x^2-(\sum x)^2}\sqrt{n\sum y^2-(\sum y)^2}}.$$

Proof :

We know that covariance of X and Y is $\operatorname{cov}(x,y)$

$$=\frac{\sum(x-\bar{x})(y-\bar{y})}{n}$$

$$=\frac{1}{n}[\sum(xy-x\bar{y}-\bar{x}y+\bar{x}\bar{y}]$$

$$=\frac{1}{n}[\sum xy-\bar{y}\sum x-\bar{x}\sum y+\sum(\bar{x}\bar{y})]$$

$$=\frac{1}{n}[\sum xy-n\bar{x}\bar{y}-\cancel{n\bar{x}\bar{y}}+\cancel{n\bar{x}\bar{y}}]$$

$$=\frac{1}{n}\left[\sum xy-\cancel{n}\frac{\sum x}{n}\frac{\sum y}{\cancel{n}}\right]$$

$$=\frac{1}{n^2}[n\sum xy-(\sum x)(\sum y)]$$

and $$\sigma_x=\sqrt{\frac{1}{n}\sum x^2-\left(\frac{\sum x}{n}\right)^2}$$

(i.e) $\sigma_x = \sqrt{\dfrac{n\sum x^2 - (\sum x)^2}{n^2}}$

(i.e) $\sigma_x = \dfrac{1}{n}\sqrt{n\sum x^2 - (\sum x)^2}$

Similarly $\sigma_y = \dfrac{1}{n}\sqrt{n\sum y^2 - (\sum y)^2}$

Thus $\gamma_{xy} = \dfrac{\text{cov}(x,y)}{\sigma_x \sigma_y}$

(i.e) $\gamma_{xy} = \dfrac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$

This prove the theorem.

Theorem 2. 2 The correlation coefficient is independent of the change of origin and scale.

Proof :

We know that correlation coefficient is $\gamma_{xy} = \dfrac{\sum(x-\overline{x})(y-y)}{n\sigma_x \sigma_y}$.

Let $u_i = \dfrac{x_i - A}{h}$ and $v_i = \dfrac{y_i - B}{k}$ where $h, k > 0$

Now $u_i = \dfrac{x_i - A}{h}$ and $v_i = \dfrac{y_i - B}{k}$

Thus $x_i = A + hu_i$ and $y_i = B + kv_i$

Hence $\overline{x} = A + h\overline{u}$ and $\overline{y} = B + k\overline{v}$

$\therefore\ x_i - \overline{x} = h(u_i - \overline{u})$ and $y_i - \overline{y} = h(v_i - \overline{v})$

and $\sigma_x = h\sigma_u$, $\sigma_y = k\sigma_v$.

Hence $\gamma_{xy} = \dfrac{\text{cov}(x,y)}{\sigma_x \sigma_y}$

$= \dfrac{\sum(x-\overline{x})(y-\overline{y})}{n\sigma_x \sigma_y}$

$= \dfrac{\sum \cancel{h}(u-\overline{u})\cancel{k}(v-\overline{v})}{n\cancel{h}\sigma_u \cancel{k}\sigma_v}$

$= \dfrac{\sum(u-\overline{u})(v-\overline{v})}{n\sigma_u \sigma_v}$

$= \gamma_{uv}$

Space for Hint

Space for Hint

Thus $\gamma_{xy} = \gamma_{uv}$

(i.e) the correlation coefficient is independent of the change of origin and scale.

This proves the theorem.

Theorem 2. 3 $-1 \le \gamma \le 1$

Proof : $\gamma_{xy} = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{n\sigma_x \sigma_y}$

$$= \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n\sqrt{\frac{1}{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}(y_i-\bar{y})^2}}$$

$$= \frac{\Sigma(a_i b_i)}{\sqrt{\Sigma a_i^2}\sqrt{\Sigma b_i^2}}$$

(i.e) $\gamma_{xy}^2 = \dfrac{(\Sigma a_i b_i)^2}{(\Sigma a_i^2)(\Sigma b_i^2)}$ -- (2.1)

We know that from Schwartz inequality $(\Sigma a_i b_i)^2 \le (\Sigma a_i^2)(\Sigma b_i^2)$

Thus (2.1) $\Rightarrow \gamma_{xy}^2 \le 1$

(i.e) $|\gamma_{xy}| \le 1$

(i.e) $-1 \le \gamma \le 1$

This proves the theorem.

Note :

(i) The correlation γ is said to be perfectly positive if $\gamma = 1$

(ii) The correlation γ is said to be perfectly negative if $\gamma = -1$

Theorem 2. 4 $\gamma_{xy} = \dfrac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x \sigma_y}$

Proof : We know that $\sigma_{x-y}^2 = \frac{1}{n}\Sigma\left[(x_i - y_i) - (\bar{x}-\bar{y})\right]^2$

$$= \frac{1}{n}\Sigma\left[(x_i-\bar{x}) - (y_i-\bar{y})\right]^2$$

$$= \frac{1}{n}\left[\Sigma(x_i-\bar{x})^2 - 2\Sigma(x_i-\bar{x})(y_i-\bar{y}) + \Sigma(y_i-\bar{y})^2\right]$$

$$= \sigma_x^2 - 2\gamma_{xy}\sigma_x\sigma_y + \sigma_y^2$$

Space for Hint

$$\therefore \quad \gamma_{xy} = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x \sigma_y}.$$

This proves the theorem.

Example 2. 1 :

Find the correlation coefficient to the following data :

x	:	10	12	18	24	23	27
y	:	13	18	12	25	30	10

Solution :

Given that

x	:	10	12	18	24	23	27
y	:	13	18	12	25	30	10

We know that $\gamma_{xy} = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{n\sigma_x \sigma_y}$.

	x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
	10	13	-9	81	-5	25	45
	12	18	-7	49	0	0	0
	18	12	-1	1	-6	36	6
	24	25	5	25	7	49	35
	23	30	4	16	12	144	48
	27	10	8	64	-8	64	-64
total	114	108	0	236	0	318	70

Now $\bar{x} = \dfrac{\Sigma x}{n}$

$= \dfrac{144}{6}$

$= 19$

Space for Hint

and $\bar{y} = \frac{\sum y}{n}$

$= \frac{108}{6}$

$= 18$

and $\sigma_x = \sqrt{\frac{1}{n}\Sigma(x-\bar{x})^2}$

$= \frac{236}{6}$

$= 6.272$

and $\sigma_y = \sqrt{\frac{1}{n}\Sigma(y-\bar{y})^2}$

$= \frac{318}{6}$

$= 7.280$

Now $\gamma_{xy} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n\sigma_x\sigma_y}$

$= \frac{70}{6(6.272)(7.280)}$

$= 0.256$

Thus the correlation coefficient between X and Y is 0.256

Example 2. 2 :

Find the correlation coefficient to the following data :

Age of husband	:	23	27	28	29	30	31	33	35	36	39
Age of wife	:	18	22	23	24	25	26	28	29	30	32

Solution :

Given that

Age of husband	:	23	27	28	29	30	31	33	35	36	39
Age of wife	:	18	22	23	24	25	26	28	29	30	32

Space for Hint

We know that $\gamma_{xy} = \dfrac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$.

	x	y	x^2	y^2	xy
	23	18	529	324	414
	27	22	729	484	594
	28	23	784	529	644
	29	24	841	576	696
	30	25	900	625	750
	31	26	961	676	806
	33	28	1089	784	924
	35	29	1225	841	1015
	36	30	1296	900	1080
	39	32	1521	1024	1248
Total	311	257	9875	6763	8171

Now $\gamma_{xy} = \dfrac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$

$= \dfrac{10(8171) - (311)(257)}{\sqrt{10(9875) - (311)^2}\sqrt{10(6763) - (257)^2}}$

$= \dfrac{1783}{(45.044)(39.762)}$

$= 0.996$

Thus the correlation coefficient between X and Y is 0.996

Space for Hint

Example 2. 3 :

Find the correlation coefficient to the following data :

Marks in Mathematics	:	65	66	67	67	68	69	70	72
Marks in Statistics	:	67	68	65	68	72	72	69	71

Solution :

Given that

Marks in Mathematics	:	65	66	67	67	68	69	70	72
Marks in Statistics	:	67	68	65	68	72	72	69	71

We know that $\gamma_{xy} = \dfrac{n\sum uv - (\sum u)(\sum v)}{\sqrt{n\sum u^2 - (\sum u)^2}\sqrt{n\sum v^2 - (\sum v)^2}}$ where $u = x - A$ and $v = y - B$.

Choose $A = 67$ and $B = 68$

	x	y	$u = x - 67$	$v = y - 68$	u^2	v^2	uv
	65	67	-2	-1	4	1	2
	66	68	-1	0	1	0	0
	67	65	0	-3	0	9	0
	67	68	0	0	0	0	0
	68	72	1	4	1	16	4
	69	72	2	4	4	16	8
	70	69	3	1	9	1	3
	72	71	5	3	25	9	15
Total	544	552	8	8	44	52	32

Now $\gamma_{xy} = \dfrac{n\sum uv - (\sum u)(\sum v)}{\sqrt{n\sum u^2 - (\sum u)^2}\sqrt{n\sum v^2 - (\sum v)^2}}$

$= \dfrac{8(32) - (8)(8)}{\sqrt{8(44) - (8)^2}\sqrt{8(52) - (8)^2}}$

$= \dfrac{192}{(16.971)(18.762)}$

$= 0.603$

Thus the correlation coefficient between X and Y is 0.603

Check Your Progress :

(1) Find the correlation coefficient to the following data :

X	:	300	350	400	450	500	550	600	650	700
Y	:	800	900	1000	1100	1200	1300	1400	1500	1600

(2) Find the correlation coefficient to the following data :

Father's height	:	67	68	64	67	72	70	70	69	70
Son's height	:	65	66	67	68	68	69	71	72	72

Example 2. 4 :

If $\sum x = 71$, $\sum y = 70$, $\sum x^2 = 555$, $\sum y^2 = 526$, $\sum xy = 527$ and $n = 100$, then find γ_{xy}.

Solution :

Given that $\sum x = 71$, $\sum y = 70$, $\sum x^2 = 555$, $\sum y^2 = 526$, $\sum xy = 527$ and $n = 100$.

We know that $\gamma_{xy} = \dfrac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$.

$\therefore \gamma_{xy} = \dfrac{100(527) - (71)(70)}{\sqrt{100(555) - (71)^2}\sqrt{100(526) - (70)^2}}$

$= \dfrac{47730}{(224.63)(218.40)}$

$= 0.9729$

Space for Hint

Space for Hint

Example 2. 5 :

Given $n = 1000$, $\bar{x} = 65$, $\bar{y} = 83$, $\sigma_x = 4.5$, $\sigma y = 3.6$ and the sum of the products of the deviations from the mean of x and y is 4800. Find the correlation coefficient between x and y.

Solution :

Given that $n = 1000$, $\bar{x} = 65$, $\bar{y} = 83$, $\sigma_x = 4.5$, $\sigma y = 3.6$ and

$\Sigma(x - \bar{x})(y - \bar{y}) = 4800$

We know that $\gamma_{xy} = \dfrac{\Sigma(x - \bar{x})(y - \bar{y})}{n\sigma_x \sigma_y}$.

$$\therefore \gamma_{xy} = \frac{4800}{1000(4.5)(3.6)}$$

$$= 0.2963$$

Example 2. 6 :

If $z = ax + by$ and γ is the correlation coefficient between x and y show that $\sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\gamma\sigma_x\sigma_y$. Hence deduce that

$$\gamma = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}.$$

Proof :

Given that $z = ax + by$

$\therefore \bar{z} = a\bar{x} + b\bar{y}$

Now $z - \bar{z} = a(x - \bar{x}) + b(y - \bar{y})$

and $(z - \bar{z})^2 = a^2(x - \bar{x})^2 + b^2(y - \bar{y})^2 + 2ab(x - \bar{x})(y - \bar{y})$

We know that $\gamma_{xy} = \gamma$

$$= \frac{\Sigma(x - \bar{x})(y - \bar{y})}{n\sigma_x\sigma_y}$$

$$\therefore \frac{1}{n}\Sigma(x - \bar{x})(y - \bar{y}) = \gamma\sigma_x\sigma_y$$

Thus $\sigma_z^2 = \dfrac{1}{n}\Sigma(z - \bar{z})^2$

$$= \frac{1}{n}\Sigma\left(a^2(x-\overline{x})^2 + b^2(y-\overline{y})^2 + 2ab(x-\overline{x})(y-\overline{y})\right)$$

$$= a^2\frac{1}{n}\Sigma(x-\overline{x})^2 + b^2\frac{1}{n}\Sigma(y-\overline{y})^2 + 2ab\frac{1}{n}\Sigma(x-\overline{x})(y-\overline{y})$$

$$= a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\gamma\sigma_x\sigma_y \text{ ------------------------------ } (2.2)$$

Deduction :

Putting $a = b = 1$ in (2.2), we get,

$$\sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2 - \gamma 2\sigma_x\sigma_y$$

Thus $\gamma = \dfrac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$.

This proves the problem.

Example 2. 7 :

If x and y are discrete variables and if $\sigma_x^2 = \sigma_y^2 = \sigma$ and $\text{cov}(x,y) = \frac{1}{2}\sigma^2$

find (i) σ_{2x-3y} and (ii) $\sigma_{2x+3,2y-3}$

Solution : (i)

Let $u = 2x - 3y$

$\therefore\ \overline{u} = 2\overline{x} - 3\overline{y}$

and $u - \overline{u} = 2(x-\overline{x}) - 3(y-\overline{y})$.

Hence $(u-\overline{u})^2 = 4(x-\overline{x})^2 + 9(y-\overline{y})^2 - 12(x-\overline{x})(y-\overline{y})$

$\therefore\ \sigma_u^2 = 4\sigma_x^2 + 9\sigma_y^2 - 12\text{cov}(x,y)$

$= 4\sigma^2 + 9\sigma^2 - 12 \times \frac{1}{2} \times \sigma^2$

$= 7\sigma^2$

Thus $\sigma_u = \sqrt{7}\sigma$

(i.e) $\sigma_{2x-3y} = \sqrt{7}\sigma$

Solution : (ii)

Let $u = 2x + 3$ and $v = 2y - 3$

$\therefore\ \overline{u} = 2\overline{x} + 3$ and $\overline{v} = 2\overline{y} - 3$

and $(u-\overline{u})(v-\overline{v}) = 4(x-\overline{x})(y-\overline{y})$

(i.e) $\text{cov}(u,v) = 4\,\text{cov}(x,y)$

Space for Hint

Space for Hint

$$= 4 \times \frac{1}{2}\sigma^2$$

$$= 2\sigma^2$$

and $\sigma_u^2 = 4\sigma_x^2, \quad \sigma_v^2 = 4\sigma_y^2$

(i.e) $\sigma_u^2 = 4\sigma^2, \quad \sigma_v^2 = 4\sigma^2$

Thus $\gamma_{uv} = \dfrac{\text{cov}(u,v)}{\sigma_u \sigma_v}$

$$= \frac{2\sigma^2}{4\sigma^2}$$

$$= \frac{1}{2}.$$

Example 2. 8 :

A computer while calculating the correlation coefficient between two variables x and y obtained in the following constants.

$n = 25, \quad \sum x = 125, \quad \sum y = 100, \quad \sum x^2 = 650, \quad \sum y^2 = 460$ and $\sum xy = 508$. It was later found that at the time of checking that operator had copied down two pairs of observations (x_i, y_i) as $(6,14)$ and $(8,6)$ instead of the correct values $(8,12)$ and $(6,8)$. Obtain the correct value of the correlation coefficient between x and y.

Solution :

Given that

Wrong (x_i, y_i)	Correct(x_i, y_i)
(6,14)	(8,12)
(8,6)	(6,8)

Correct $\sum x$ = wrong $\sum x$ – (sum of wrong items) + (sum of correct items)

$= 125 - (6 + 8) + (8 + 6)$

$= 125$

Correct $\sum y$ = wrong $\sum y$ – (sum of wrong items) + (sum of correct items)

$= 100 - (14 + 6) + (12 + 8)$

$= 100$

Space for Hint

Correct $\sum x^2$ = wrong $\sum x^2$ – (sum of squares of wrong items)

+ (sum of squares of correct items)

$= 650 - (6^2 + 8^2) + (8^2 + 6^2)$

$= 650$

Correct $\sum y^2$ = wrong $\sum y^2$ – (sum of squares of wrong items)

+ (sum of squares of correct items)

$= 460 - (14^2 + 6^2) + (12^2 + 8^2)$

$= 436$

Correct $\sum xy$ = wrong $\sum xy$ – (sum of product of wrong items)

+ (sum of product of correct items)

$= 508 - (6 \times 14 + 8 \times 6) + (8 \times 12 + 6 \times 8)$

$= 520$

Thus correct correlation coefficient = γ_{xy}

$$= \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$$

$$= \frac{25(520) - (125)(100)}{\sqrt{25(650) - (125)^2}\sqrt{25(436) - (100)^2}}$$

$$= \frac{500}{25 \times 30}$$

$= 0.667$

Hence the correlation coefficient is 0.667.

Check Your Progress :

(1) Find the correlation coefficient for the data.

x :	120	110	120	119	140	125	127	119	140	160
y :	240	250	260	266	232	245	255	267	268	239

(2) Find the correlation coefficient for the data.

x :	300	350	400	450	500	550	600	650	700
y :	800	900	1000	1100	1200	1300	1400	1500	1600

(answers : (1) -0.373, (2) 1)

Space for Hint

2. 2 Bivariate Correlation

When two variables are given in the bivariate frequency table, then we can able to find the correlation coefficient.

Example 2. 9 :

Find the correlation coefficient to following bivariate frequency data.

ages of mothers	ages of daughters					total
	20 – 30	30 – 40	40 – 50	50 – 60	60 - 70	
15-25	5	9	3	-	-	17
25-35	-	10	25	2	-	37
35-45	-	1	12	2	-	15
45-55	-	-	4	16	5	25
55-65	-	-	-	4	2	6
total	5	20	44	24	7	100

Solution :

We shall find the values of $\Sigma f_i u_i$, $\Sigma f_i v_i$, $\Sigma f_i u_i^2$, $\Sigma f_i v_i^2$ and $\Sigma f_i u_i v_j$.

We know that $\gamma_{xy} = \dfrac{N\Sigma f_i u_i v_j - (\Sigma f_i u_i)(\Sigma f_i v_j)}{\sqrt{N\Sigma f_i u_i^2 - (\Sigma f_i u_i)^2}\sqrt{N\Sigma f_i v_j^2 - (\Sigma f_i v_j)^2}}$.

(Calculations are shown the following table)

$$\therefore \gamma_{xy} = \frac{100(88) - (8)(-34)}{\sqrt{100(92) - (8)^2}\sqrt{100(154) - (-34)^2}}$$

$$= \frac{9072}{95.582 \times 119.35}$$

$$= 0.7953$$

	u_i	-2	-1	0	1	2				
v_j		25	35	45	55	65	f_j	f_jv_j	$f_jv_j^2$	$f_jv_ju_i$
-2	20	(20) 5	(18) 9	(0) 3	- -	- -	17	-34	68	38
-1	30	- -	(10) 10	(0) 25	(-2) 2	- -	37	-37	37	8
0	40	- -	(1) 1	() 12	(-2) 2	- -	15	0	0	
1	50	- -	- -	(0) 4	(16) 16	(10) 5	25	25	25	26
2	60	- -	- -	- -	(8) 4	(8) 2	6	12	24	16
f_i		5	20	44	24	7	100	-34	154	88
f_iu_i		-10	-20	0	24	14	8	check		
$f_iu_i^2$		20	20	0	24	28	92			
$f_iu_iv_j$		20	28	0	22	18	88			

Example 2. 10 :

Find the correlation coefficient to following bivariate frequency data.

ages of mothers	ages of daughters					total
	5-10	10-15	15-20	20-25	25-30	
15-25	6	3	-	-	-	9
25-35	3	10	10	-	-	23
35-45	-	10	15	7	-	32
45-55	-	-	7	10	4	21
55-65	-	-	-	4	5	9
total	9	29	32	21	9	100

Space for Hint

Space for Hint

Solution :

We shall find the values of $\sum f_i u_i$, $\sum f_i v_i$, $\sum f_i u_i^2$, $\sum f_i v_i^2$ and $\sum f_i u_i v_j$.

	u_i	-2	-1	0	1	2				
v_j		7.5	12.5	17.5	22.5	27.5	f_j	$f_j v_j$	$f_j v_j^2$	$f_j v_j u_i$
-2	20	(24) 6	(6) 3	(0) -	- -	- -	9	-18	36	30
-1	30	- 3	(16) 16	(0) 10	- -	- -	29	-29	29	16
0	40	- -	(10) 10	(0) 15	(-7) 7	- -	32	0	0	
1	50	- -	- -	(0) 7	(10) 10	(8) 4	21	21	21	18
2	60	- -	- -	- -	(8) 4	(20) 5	9	18	36	28
f_i		9	29	32	21	9	100	-8	122	92
$f_i u_i$		-18	-29	0	21	18	-8			
$fi\ u_i^2$		36	29	0	21	36	122			
$f_i u_i v_j$		24	22	0	18	28	92			

We know that $\gamma_{xy} = \dfrac{N\sum f_i u_i v_j - (\sum f_i u_i)(\sum f_i v_j)}{\sqrt{N\sum f_i u_i^{\,2} - (\sum f_i u_i)^2}\ \sqrt{N\sum f_i v_j^{\,2} - (\sum f_i v_j)^2}}$.

$$\therefore\ \gamma_{xy} = \frac{100(92) - (-8)(-8)}{\sqrt{100(122) - (-8)^2}\ \sqrt{100(122) - (-8)^2}}$$

$$= \frac{9136}{110.164 \times 110.164}$$

$$= 0.753$$

Check Your Progress :

(1) Find the correlation coefficient to following bivariate frequency data.

Marks	ages in years					Total
	18	19	20	21	22	
20 – 25	3	2	-	-	-	5
15 – 20	-	5	4	-	-	9
10 – 15	-	-	7	10	-	17
5 – 10	-	-	-	3	2	5
0 – 5	-	-	-	3	1	4
Total	3	7	11	16	3	100

(2) The following table gives the marks in Algebra and Statistics got by 50 students in an internal assessment test. Calculate the correlation coefficient.

Marks	ages in years					Total
	18	19	20	21	22	
20 – 25	3	2	-	-	-	5
15 – 20	-	5	4	-	-	9
10 – 15	-	-	7	10	-	17
5 – 10	-	-	-	3	2	5
0 – 5	-	-	-	3	1	4
Total	3	7	11	16	3	100

(Answer (1) – 0.837, (2) 0.28)

Space for Hint

Space for Hint

2. 3 Rank Correlation Coefficient

De ... **d is defined**

as $\rho = 1 - \dfrac{6\sum(x-y)^2}{n(n^2-1)}$.

Note : If two or more individuals get the same rank then the rank correlation formula is modified as $\rho = 1 - \dfrac{6\left(\sum(x-y)^2 + CF\right)}{n(n^2-1)}$ where

CF = correction factor and it is obtained from

$$CF = \frac{1}{12}m_1(m_1^2-1) + \frac{1}{12}m_2(m_2^2-1) + \cdots,$$

here $m_1, m_2, \cdots$ are the number of times that ranks be repeated.

Example 2. 11 :

Find the rank correlation coefficient for the following data.

x :	5	2	8	1	4	6	3	7
y :	4	5	7	3	2	8	1	6

Solution :

Given that

x :	5	2	8	1	4	6	3	7
y :	4	5	7	3	2	8	1	6

We know that rank correlation = ρ

$$= 1 - \frac{6\sum(x-y)^2}{n(n^2-1)}$$

x	y	$x-y$	$(x-y)^2$
5	4	1	1
2	5	-3	9
8	7	1	1
1	3	-2	4
4	2	2	4
6	8	-2	4
3	1	2	4
7	6	1	1
	total		28

Thus $\rho = 1 - \dfrac{6 \times 28}{8 \times (8^2 - 1)}$

$= 1 - \dfrac{168}{504}$

$= 0.667$

(i.e) the rank correlation coefficient is 0.667.

Example 2. 12 :

Find the rank correlation coefficient for the following data.

X :	78	65	36	98	25	75	82	90	62	39
Y :	84	53	51	91	60	68	62	86	58	47

Solution :

Given that

X :	78	65	36	98	25	75	82	90	62	39
Y :	84	53	51	91	60	68	62	86	58	47

We know that rank correlation = ρ

$= 1 - \dfrac{6\sum(x-y)^2}{n(n^2-1)}$

Space for Hint

Space for Hint

X	Y	Rank of $X(x)$	Rank of $Y(y)$	$x-y$	$(x-y)^2$
78	84	4	3	1	1
65	53	6	8	-2	4
36	51	9	9	0	0
98	91	1	1	0	0
25	60	10	6	4	16
75	68	5	4	1	1
82	62	3	5	-2	4
90	86	2	2	0	0
62	58	7	7	0	0
39	47	8	10	-2	4
		total			30

Thus $\rho = 1 - \frac{6 \times 30}{10 \times (10^2 - 1)}$

$= 1 - \frac{180}{990}$

$= 0.818$

(i.e) the rank correlation coefficient is 0.818.

Example 2. 13 :

Find the rank correlation coefficient for the following data.

X :	48	58	38	28	60	38	54	60	40	56
Y :	74	70	32	52	46	54	38	40	32	22

Solution :

Given that

X :	48	58	38	28	60	38	54	60	40	56
Y :	74	70	32	52	46	54	38	40	32	22

We know that rank correlation = ρ

$$= 1 - \frac{6(\sum(x-y)^2 + CF)}{n(n^2-1)}$$

where $CF = \frac{1}{12}m_1(m_1^2 - 1) + \frac{1}{12}m_2(m_2^2 - 1) + \cdots,$

X	Y	Rank of $X(x)$	Rank of $Y(y)$	$x-y$	$(x-y)^2$
48	74	6.0	1.0	5.00	25.00
58	70	3.0	2.0	1.00	1.00
38	32	8.5	9.5	-1.00	1.00
28	52	10.0	4.0	6.00	36.00
60	46	1.5	5.0	-3.50	12.25
38	54	8.5	3.0	5.50	30.25
54	38	5.0	8.0	-3.00	9.00
60	40	1.5	7.0	-5.50	30.25
40	32	7.0	9.5	-2.50	6.25
56	22	4.0	11.0	-7.00	49.00
22	42	11.0	6.0	5.00	25.00
total					225.00

Now $CF = \frac{1}{12}m_1(m_1^2 - 1) + \frac{1}{12}m_2(m_2^2 - 1) + \cdots$

$$= \frac{1}{12}2(2^2 - 1) + \frac{1}{12}2(2^2 - 1) + \frac{1}{12}2(2^2 - 1)$$

$$= 1.5$$

Thus $\rho = 1 - \frac{6(\sum(x-y)^2 + CF)}{n(n^2-1)}$

$$= 1 - \frac{6 \times (225 + 1.5)}{11(11^2 - 1)}$$

$$= 1 - \frac{1359}{11(11^2 - 1)}$$

$$= 1 - \frac{1359}{1320}$$

$$= -0.0295$$

Space for Hint

Space for Hint

(i.e) the rank correlation coefficient is – 0.0295.

Example 2. 14 :

Find the rank correlation coefficient for the following data.

X :	115	109	112	87	98	98	120	100	98	118
Y :	75	73	85	70	76	65	82	73	68	80

Solution :

Given that

X :	115	109	112	87	98	98	120	100	98	118
Y :	75	73	85	70	76	65	82	73	68	80

We know that rank correlation = ρ

$$= 1 - \frac{6(\sum(x-y)^2 + CF)}{n(n^2-1)}$$

where $CF = \frac{1}{12}m_1(m_1^2 - 1) + \frac{1}{12}m_2(m_2^2 - 1) + \cdots,$

X	Y	Rank of $X(x)$	Rank of $Y(y)$	$x-y$	$(x-y)^2$
115	75	3	5.0	-2.00	4.00
109	73	5	6.5	-1.50	2.25
112	85	4	1.0	3.00	9.00
87	70	10	8.0	2.00	4.00
98	76	8	4.0	4.00	16.00
98	65	8	10.0	-2.00	4.00
120	82	1	2.0	-1.00	1.00
100	73	6	6.5	-0.50	0.25
98	68	8	9.0	-1.00	1.00
118	80	2	3.0	-1.00	1.00
total					42.50

Now $CF = \frac{1}{12}m_1(m_1^2 - 1) + \frac{1}{12}m_2(m_2^2 - 1) + \cdots$

$$= \frac{1}{12}3(3^2 - 1) + \frac{1}{12}2(2^2 - 1)$$

$$= 2.5$$

Thus $\rho = 1 - \frac{6(\sum(x-y)^2 + CF)}{n(n^2-1)}$

$= 1 - \frac{6 \times (42.5+2.5)}{10(10^2-1)}$

$= 1 - \frac{42.50}{990}$

$= 0.7273$

(i.e) the rank correlation coefficient is 0.7273.

Example 2. 15 :

Ten competitors in a beauty contest were ranked by three judges in the following order :

Judge I :	1	6	5	10	3	2	4	9	7	8
Judge II :	3	5	8	4	7	10	2	1	6	9
Judge III :	6	4	9	8	1	2	3	10	5	7

Solution :

Step 1 : To find the rank correlation coefficient between Judges I and II.

Now

Judge I (x) :	1	6	5	10	3	2	4	9	7	8
Judge II (y) :	3	5	8	4	7	10	2	1	6	9

Now we shall find $\sum(x-y)^2$

x	y	$x-y$	$(x-y)^2$
1	3	-2	4
6	5	1	1
5	8	-3	9
10	4	6	36
3	7	-4	16
2	10	-8	64
4	2	2	4
9	1	8	64
7	6	1	1
8	9	-1	1
Total			195

Space for Hint

Space for Hint

We know that rank correlation = ρ

$$= 1 - \frac{6\sum(x-y)^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 195}{10 \times (10^2 - 1)}$$

$$= 1 - \frac{1170}{990}$$

$$= -0.18182$$

(i.e) the rank correlation coefficient is -0.182.

Step 2 : To find the rank correlation coefficient between Judges II and III.

Now

Judge II (y) :	3	5	8	4	7	10	2	1	6	9
Judge III (z) :	6	4	9	8	1	2	3	10	5	7

Now we shall find $\sum(y-z)^2$

y	z	$y-z$	$(y-z)^2$
3	6	-3	9
5	4	1	1
8	9	-1	1
4	8	-4	16
7	1	6	36
10	2	8	64
2	3	-1	1
1	10	-9	81
6	5	1	1
9	7	2	4
total			204

We know that rank correlation = ρ_{yz}

$$= 1 - \frac{6\sum(x-y)^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 204}{10 \times (10^2 - 1)}$$

$$= 1 - \frac{1224}{990}$$

$$= -0.23636$$

(i.e) the rank correlation coefficient is -0.236.

Step 3 : To find the rank correlation coefficient between Judges III and I.

Now

Judge III (z) :	6	4	9	8	1	2	3	10	5	7
Judge I (x) :	1	6	5	10	3	2	4	9	7	8

Now we shall find $\sum(y-z)^2$

y	z	$y-z$	$(y-z)^2$
6	1	5	25
4	6	-2	4
9	5	4	16
8	10	-2	4
1	3	-2	4
2	2	0	0
3	4	-1	1
10	9	1	1
5	7	-2	4
7	8	-1	1
total			31

We know that rank correlation = ρ_{zx}

$$= 1 - \frac{6\sum(x-y)^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 31}{10 \times (10^2 - 1)}$$

$$= 1 - \frac{186}{990}$$

$$= 0.812121$$

(i.e) the rank correlation coefficient is 0.812.

Space for Hint

Space for Hint

Step 4 : Clearly from step 3, the judge I and judge III have a positive rank correlation and there these judges have nearest taste in beauty.

Check Your Progress :

(1) Find the rank correlation coefficient to the following data.

X :	92	89	87	86	83	77	71	63	53	50
Y :	86	83	91	77	68	85	52	82	37	57

(2) Find the rank correlation coefficient to the following data.

X :	30	50	25	30	60	70	80	65	75	98
Y :	50	60	20	40	70	40	90	60	40	80

(3) Find the rank correlation coefficient to the following data.

X :	48	33	40	9	16	16	65	24	16	57
Y :	13	13	24	6	15	4	20	9	6	19

(answers : (1) 0.73, (2) 0.61, (3) 0.73)

2. 4 Regression

In a functional relationship, if we know the value of one variable, the value of the other variable is determined exactly. In a statistical relationship, the situation is slightly different. That is, we cannot determine the value of one variable from that of the other variable. By regression, we can able to find the approximate value of one variable from the other. Hence regression is the measure of the average relationship between two or more variables in terms of the original units of the data.

Space for Hint

A regression equation describes the rule to be followed for determining the predicted value of one variable from a given value of the other variable. Thus the regression equation of y on x is used to predict the values of y from the given values of x and the regression equation of x on y is used to obtain the predicted value of x from the given value of y.

Definition : The regression equation of x on y is $x-\overline{x}=\dfrac{\gamma\sigma_x}{\sigma_y}(y-\overline{y})$

or $x-\overline{x}=b_{xy}(y-\overline{y})$ where $b_{xy}=\dfrac{\gamma\sigma_x}{\sigma_y}$.

Similarly the regression equation of y on x is $y-\overline{y}=\dfrac{\gamma\sigma_y}{\sigma_x}(x-\overline{x})$

or $y-\overline{y}=b_{yx}(x-\overline{x})$ where $b_{xy}=\dfrac{\gamma\sigma_x}{\sigma_y}$.

Note :

(1) b_{xy} is called regression coefficient of x on y and b_{yx} is called y on x.

(2) $b_{xy}\cdot b_{yx}=\gamma^2$

(3) $\gamma=\pm\sqrt{b_{xy}\cdot b_{yx}}$ and the sign of the correlation coefficient is the same as that of the regression coefficients.

Theorem 2. 5

If one of the regression coefficients is greater than unity then the other is less than unity.

Proof : We know that $b_{xy}\cdot b_{yx}=\gamma^2$ and $-1\le\gamma\le1$

$\therefore\ \gamma^2\le1$

(i.e) $b_{xy}\cdot b_{yx}\le1$

Thus if $b_{xy}>1\Rightarrow b_{yx}<1$.

Hence one of the regression coefficients is greater than unity then the other is less than unity.

This proves the theorem.

Space for Hint

Theorem 2. 6

Arithmetic mean of the regression coefficients is greater than or equal to the correlation coefficient.

Proof : We know that the regression coefficients are b_{xy} and b_{yx}.

To prove that $\frac{1}{2}(b_{xy}+b_{yx}) \geq \gamma$

Now $\frac{1}{2}(b_{xy}+b_{yx}) \geq \gamma$

$\Leftrightarrow \; b_{xy}+b_{yx} \geq 2\gamma$

$\Leftrightarrow \; \frac{\gamma\sigma_x}{\sigma_y}+\frac{\gamma\sigma_y}{\sigma_x} \geq 2\gamma$

$\Leftrightarrow \; \frac{\sigma_x}{\sigma_y}+\frac{\sigma_y}{\sigma_x} \geq 2$

$\Leftrightarrow \; \sigma_x^2+\sigma_y^2 \geq 2\sigma_x\sigma_y$

$\Leftrightarrow \; \sigma_x^2+\sigma_y^2-2\sigma_x\sigma_y \geq 0$

$\Leftrightarrow \; (\sigma_x-\sigma_y)^2 \geq 0$

which always true.

Thus $\frac{1}{2}(b_{xy}+b_{yx}) \geq \gamma$

This proves the theorem.

Theorem 2. 7

Regression coefficients are independent of the change of origin but dependent on change of scale.

Proof : Let $u_i = \frac{x_i - A}{h}$ and $v_j = \frac{y_j - B}{k}$.

Now $u_i = \frac{x_i - A}{h} \Rightarrow x_i = A + hu_i$ and

$v_j = \frac{y_j - B}{k} \Rightarrow y_j = B + kv_j$

We know that $\sigma_x = h\sigma_u$, $\sigma_y = k\sigma_v$ and $\gamma_{xy} = \gamma_{uv}$

Space for Hint

Now $b_{xy} = \gamma_{xy}\dfrac{\sigma_x}{\sigma_y}$

$$= \gamma_{uv}\frac{h\sigma_u}{k\sigma_v}$$

$$= \frac{h}{k}b_{vu} \text{ -- (2.3)}$$

Similarly $b_{yx} = \dfrac{k}{h}b_{vu}$ -- (2.4)

From (2.3) and (2.4) it is clear that b_{xy} and b_{yx} depend upon h and k but not on origin A and B.

Thus Regression coefficients are independent of the change of origin but dependent on change of scale.

Theorem 2. 8

The angle between two regression lines is $\tan^{-1}\left[\left(\dfrac{\gamma^2 \sim 1}{\gamma}\right)\left(\dfrac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\right)\right]$.

Proof :

We know that the equation of the regression lines are $x-\overline{x} = b_{xy}(y-\overline{y})$ and $y-\overline{y} = b_{yx}(x-\overline{x})$.

Now the slopes of the regression lines are $\dfrac{1}{b_{xy}}$ and b_{yx}.

Let θ be the angle between the two regression lines.

$$\therefore \tan\theta = \frac{m_1 \sim m_2}{1+m_1 m_2}$$

$$= \frac{\frac{1}{b_{xy}} \sim b_{yx}}{1+\frac{1}{b_{xy}}\cdot b_{yx}}$$

$$= \frac{1 \sim b_{xy}\cdot b_{yx}}{b_{xy}+b_{yx}}$$

$$= \frac{1 \sim \gamma^2}{\frac{\gamma\sigma_x}{\sigma_y}+\frac{\gamma\sigma_y}{\sigma_x}}$$

Space for Hint

$$= \left(\frac{\gamma^2 \sim 1}{\gamma}\right)\left(\frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}\right)$$

Thus $\theta = \tan^{-1}\left[\left(\frac{\gamma^2 \sim 1}{\gamma}\right)\left(\frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}\right)\right]$

(i.e) the angle between the twp regression lines is $\tan^{-1}\left[\left(\frac{\gamma^2 \sim 1}{\gamma}\right)\left(\frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}\right)\right]$.

This proves the theorem.

Note :

(1) If the two variables are uncorrelated then $\gamma = 0$

$\therefore \ \theta = \tan^{-1}(\infty)$

$\Rightarrow \theta = \frac{\pi}{2}$

$\Rightarrow$ the regression lines are perpendicular to each other.

(2) If the variables are perfectly positively correlated or perfectly negatively correlated then $\gamma = \pm 1$

$\therefore \ \theta = \tan^{-1}(0)$

$\Rightarrow \theta = 0$ or π

$\Rightarrow$ the regression lines are parallel to each other.

Example 2. 16 :

The following table shows the ages x and blood pressure y of 12 persons.

(i) Find the correlation coefficient between x and y.

(ii) Estimate the regression equations.

(iii) Estimate the blood pressure of a person whose age is 46 years.

age	blood pressure	age	blood pressure
56	147	55	150
42	125	49	145
72	160	38	115
36	118	42	140
63	149	68	152
47	128	60	155

Space for Hint

Solution :

x	y	$u=x-55$	u^2	$v=y-140$	v^2	uv
56	147	1	1	7	49	7
42	125	-13	169	-15	225	195
72	160	17	289	20	400	340
36	118	-19	361	-22	484	418
63	149	8	64	9	81	72
47	128	-8	64	-12	144	96
55	150	0	0	10	100	0
49	145	-6	36	5	25	-30
38	115	-17	289	-25	625	425
42	140	-13	169	0	0	0
68	152	13	169	12	144	156
60	155	5	25	15	225	75
Total		-32	1636	4	2502	1754

Step 1 : To find $\overline{x}$.

We know that $\overline{x} = A + \frac{\sum u}{n}$

(i.e) $\overline{x} = 55 + \frac{(-32)}{12}$

$= 52.33$

Step 2 : To find $\overline{y}$.

We know that $\overline{y} = B + \frac{\sum v}{n}$

(i.e) $\overline{y} = 140 + \frac{4}{12}$

$= 140.33$

Space for Hint

Step 3 : To find σ_x.

We know that $\sigma_x = \sqrt{\frac{1}{n}\sum u^2 - \left(\frac{\sum u}{n}\right)^2}$

(i.e) $\sigma_x = \sqrt{\frac{1}{n}\sum u^2 - \left(\frac{\sum u}{n}\right)^2}$

$= \sqrt{\frac{1636}{12} - \left(\frac{-32}{12}\right)^2}$

$= \sqrt{129.22}$

$= 11.37$

Step 4 : To find σ_y.

We know that $\sigma_y = \sqrt{\frac{1}{n}\sum v^2 - \left(\frac{\sum v}{n}\right)^2}$

(i.e) $\sigma_y = \sqrt{\frac{1}{n}\sum v^2 - \left(\frac{\sum v}{n}\right)^2}$

$= \sqrt{\frac{2502}{12} - \left(\frac{4}{12}\right)^2}$

$= \sqrt{208.39}$

$= 14.44$

Step 4 : To find correlation coefficient.

We know that $\gamma = \dfrac{n\sum uv - (\sum u)(\sum v)}{\sqrt{n\sum u^2 - (\sum u)^2}\sqrt{n\sum v^2 - (\sum v)^2}}$

Now $\gamma = \dfrac{n\sum uv - (\sum u)(\sum v)}{\sqrt{n\sum u^2 - (\sum u)^2}\sqrt{n\sum v^2 - (\sum v)^2}}$

$= \dfrac{12(1754) - (-32)(4)}{\sqrt{12(1636) - (-32)^2}\sqrt{12(2502) - (4)^2}}$

$= \dfrac{21176}{136.41 \times 173.23}$

$= 0.90$

Thus the correlation between x and y is 0.90

Step 6 : To find the regression equation of x on y.

We know that the regression equation of x on y is $x-\bar{x}=\frac{\gamma\sigma_x}{\sigma_y}(y-\bar{y})$.

(i.e) $x-52.33=\frac{0.9(11.37)}{14.44}(y-140.33)$

(i.e) $x-52.33=0.706(y-140.33)$

(i.e) $x-52.33=0.706\,y-99.494$

(i.e) $x=0.706\,y-46.742$

which is the regression equation of x on y

Step 7 : To find the regression equation of y on x.

We know that the regression equation of y on x is $y-\bar{y}=\frac{\gamma\sigma_y}{\sigma_x}(x-\bar{x})$.

(i.e) $y-140.33=\frac{0.90(14.44)}{11.37}(x-52.33)$

(i.e) $y-140.33=1.138(x-52.33)$

(i.e) $y-140.33=1.138x-59.555$

(i.e) $y=1.138x+80.778$

which is the regression equation of y on x.

Step 9 : To find the blood pressure of a person whose age is 46 years.

Put $x=46$ in the regression equation of y on x, we get,

$y=1.138(46)+80.778$

(i.e) $y=133.126$

(i.e) $y=133$ (approximately)

That is the blood pressure is 133 when the age is 46 years.

Example 2. 17 :

From the following data, obtain the two regression equations.

(i) Find the correlation coefficient between x and y.

(ii) Estimate the regression equations.

(iii) Estimate the blood pressure of a person whose age is 46 years.

Space for Hint

Space for Hint

Sales	Purchase
91	71
97	75
108	69
121	97
67	70
124	91
51	39
73	61
111	80
57	47

Solution :

x	y	$u=x-55$	u^2	$v=y-140$	v^2	uv
91	71	-6	36	1	1	-6
97	75	0	0	5	25	0
108	69	11	121	-1	1	-11
121	97	24	576	27	729	648
67	70	-30	900	0	0	0
124	91	27	729	21	441	567
51	39	-46	2116	-31	961	1426
73	61	-24	576	-9	81	216
111	80	14	196	10	100	140
57	47	-40	1600	-23	529	920
Total		-70	6850	0.0	2868	3900

Step 1 : To find $\bar{x}$.

We know that $\bar{x} = A + \frac{\sum u}{n}$

$$(\text{i.e}) \ \bar{x} = 97 + \frac{(-70)}{10}$$

$$= 90$$

Step 2 : To find $\bar{y}$.

We know that $\bar{y} = B + \frac{\sum v}{n}$

$$(\text{i.e}) \ \bar{y} = 70 + \frac{0}{10}$$

$$= 70$$

Step 3 : To find b_{xy}.

We know that $b_{xy} = \frac{n\sum uv - (\sum u)(\sum v)}{n\sum v^2 - (\sum v)^2}$

$$(\text{i.e}) \ b_{xy} = \frac{10(3900) - (-70)(0)}{10(2868) - (0)^2}$$

$$= \frac{39000}{28680}$$

$$= 1.36$$

Step 4 : To find b_{yx}.

We know that $b_{yx} = \frac{n\sum uv - (\sum u)(\sum v)}{n\sum u^2 - (\sum u)^2}$

$$(\text{i.e}) \ b_{yx} = \frac{10(3900) - (-70)(0)}{10(6850) - (0)^2}$$

$$= \frac{39000}{63600}$$

$$= 0.61$$

Step 5 : To find the regression equation of x on y.

We know that the regression equation of x on y is $x - \bar{x} = b_{xy}(y - \bar{y})$.

(i.e) $x - 90 = 1.36(y - 70)$

(i.e) $x - 90 = 1.36\,y - 95.20$

(i.e) $x = 1.36\ y - 5.2$

which is the regression equation of x on y

Space for Hint

Space for Hint

Step 6 : To find the regression equation of y on x.

We know that the regression equation of y on x is $y-\overline{y}=b_{yx}(x-\overline{x})$.

(i.e) $y-70=0.61(x-90)$

(i.e) $y-70=0.61\,x-54.90$

(i.e) $y=0.61\,x-15.1$

which is the regression equation of y on x.

Example 2. 18 :

For a bivariate data, the mean value of x is 20 and the mean value of y is 45. The regression coefficient of y on x is 4 and that of x on y is $\frac{1}{9}$. Find

(i) the coefficient of correlation.

(ii) the standard deviation of x if the standard deviation of y is 12.

(iii) the equations of regression lines.

Solution :

Given that $\overline{x}=20$, $\overline{y}=45$, $b_{yx}=4$, $b_{xy}=\frac{1}{9}$ and $\sigma_y=12$.

Step 1 : To find the correlation coefficient.

We know that $\gamma^2=b_{xy}\cdot b_{yx}$

(i.e) $\gamma^2=(4)\left(\frac{1}{9}\right)$

$\therefore\ \gamma=\pm\frac{2}{3}$

Since both b_{xy} and b_{yx} are positive, therefore $\gamma=\frac{2}{3}$.

Thus the correlation coefficient is $\gamma=\frac{2}{3}$.

Step 2 : To find standard deviation of x.

Now $b_{yx}=4$

(i.e) $\frac{\gamma\sigma_y}{\sigma_x}=4$

Space for Hint

(i.e) $\sigma_x = \left(\frac{1}{4}\right)\left(\frac{2}{3}\right)(12)$

(i.e) $\sigma_x = 2$.

Thus standard deviation of x is 2.

Step 3 : To find the regression equation of x on y

We know that $x-\bar{x} = b_{xy}(y-\bar{y})$

(i.e) $x-20 = \frac{1}{9}(y-45)$

(i.e) $x-20 = \frac{1}{9}y-5$

(i.e) $x = \frac{1}{9}y+15$

(i.e) $9x = y+135$

which is the required the regression equation of x on y.

Step 4 : To find the regression equation of y on x.

We know that $y-\bar{y} = b_{yx}(x-\bar{x})$

(i.e) $y-45 = 4(x-20)$

(i.e) $y-45 = 4x-80$

(i.e) $y = 4x-35$

which is the required the regression equation of y on x.

Example 2. 19 :

The lines of two regression equations are $x-y+5=0$ and $16x-9y+94=0$. Find the means of the variables and the variance of x if the variance of y is 16. Also find the covariance of x and y.

Solution : Given that the regression equations are

$x-y+5=0$ -- (2.5)

$16x-9y+94=0$ --- (2.6)

Solving (2.5) and (2.6), we get $x = -7$, and $y = -2$.

Thus $\bar{x} = -7$ and $\bar{y} = -2$

Space for Hint

Assume that (2.5) refers regression equation of x on y.

Now $x-y+5=0$

(i.e) $x=y-5$

$\therefore\ b_{xy}=1$

Thus form (2.6), $b_{yx}=\frac{16}{9}$

We know that $\gamma^2=b_{xy}b_{yx}$

$\therefore\ \gamma^2=1\left(\frac{16}{9}\right)>1$

Thus our assumption that (2.5) refers regression equation of x on y is wrong and hence (2.5) refers regression equation of y on x.

$\therefore\ b_{xy}=\frac{9}{16}$ and $b_{yx}=1$.

Now $\gamma^2=b_{xy}b_{yx}$

$\Rightarrow \gamma^2=\frac{9}{16}$

$\Rightarrow \gamma=\pm\frac{3}{4}$

Since both b_{xy} and b_{yx} are positive, therefore $\gamma=\frac{3}{4}$.

Now we shall find the standard deviation of x

Given that variance of y is 16

Now $b_{yx}=1$

(i.e) $\frac{\gamma\sigma_y}{\sigma_x}=1$

(i.e) $\sigma_x=\gamma\sigma_y$

(i.e) $\sigma_x=\frac{3}{4}(4)$

(i.e) $\sigma_x=3$

Now $\gamma=\frac{3}{4}$

Space for Hint

(i.e) $\frac{\operatorname{cov}(x,y)}{\sigma_x\sigma_y}=\frac{3}{4}$

(i.e) $\operatorname{cov}(x,y)=\frac{3}{4}\sigma_x\sigma_y$

(i.e) $\frac{\operatorname{cov}(x,y)}{\sigma_x\sigma_y}\frac{\sigma_y}{\sigma_x}=1$

(i.e) $\operatorname{cov}(x,y)=\sigma_x^2$

(i.e) $\operatorname{cov}(x,y)=9$

Thus the covariance between x and y is 9.

Check Your Progress

(1) The following data relate to the marks of 10 students in the two subjects Mathematics and Statistics examinations for the maximum of 50 marks.

Marks in Mathematics	25	28	30	32	35	36	38	39	42	45
Marks in Statistics	20	26	29	30	25	18	26	35	35	46

(a) obtain the two regression equations and determine

(b) the most likely Mathematics mark for the Statistics mark of 25

(c) the most likely Statistics mark for the Mathematics mark of 30.

(Answer : $y=0.905x-2.675$, $x=0.542y+19.282$, 32.83, 24.48)

(2) Two variables x and y have the regression lines $3x+2y-26=0$ and $6x+y-31=0$. Find

(a) the mean values of x and y

(b) the correlation coefficient between x and y

(c) the variance of y if the variance of x is 25

(Answer : $\bar{x}=4$, $\bar{y}=7$, $\gamma=-0.5$ and variance of y = 225)

Space for Hint

(3) In a partially destroyed laboratory record of an analysis of coefficient data the following results only are legible. Variance of x is 25, regression equations :

$3x+2y-26=0$ and $6x+y-51=0$. Find

(a) the mean values of x and y

(b) the standard deviation of x and y

(c) the correlation coefficient between x and y.

(4) From the regression lines $x=19.13-0.87$ and $y=11.64-0.50x$, find mean values of x and y, the correlation coefficient between x and y.

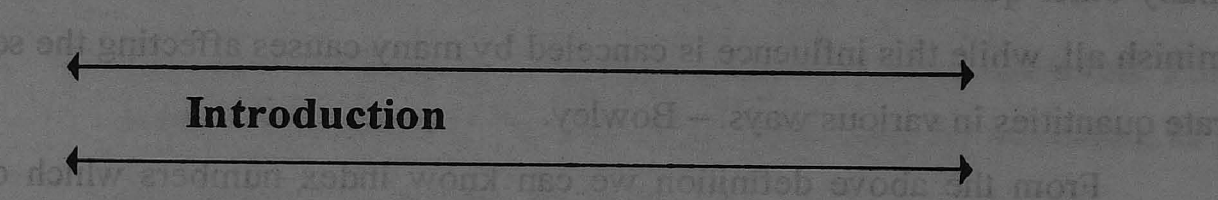
SUMMARY

In this unit we have learned the method of finding correlation coefficient between two variables, rank correlation, bivariate correlation coefficient and regression equations.

Space for Hint

UNIT 3

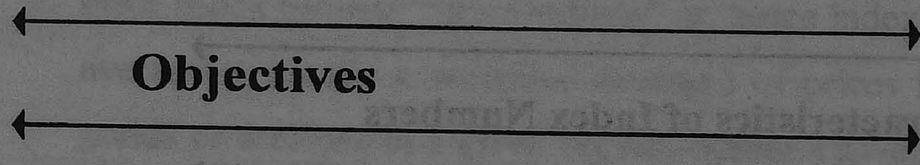
INDEX NUMBERS AND TIME SERIES



Introduction

For comparing many business and economic problems, relative numbers obtained by reducing the data are used. These relative numbers comparing any two situations comprising of same type of variables are called 'index numbers'. Index numbers are used when one is trying to compare series of numbers of vastly different size. It is a way to standardize the measurement of numbers so that they are directly comparable. Students you all would have come across this tool used in macro economics for stating inflation, consumer price index, wholesale price index, growth index, industrial out put index etc. These indexes are statistical numbers for decision making and information.

Later we shall discuss analysis of time series



Objectives

The objectives of the unit are as follows

1. To understand the various types of index numbers
2. To construct consumer price index
3. To know about the various tests of perfection for index numbers

Introduction

These index numbers are defined as follows

Index numbers are devices for mitigating deceptions caused by changes in the value of money – Marris. R

Index number represents the general level of magnitude of changes between two or more situations of a number of variables taken as a whole. – Karmel. P. H.

Space for Hint

Index numbers are used to measure the changes in some quantity which can be observed directly which we know to have a definite influence on many other qualities which we can so observe, tending to increase all or diminish all, while this influence is canceled by many causes affecting the separate quantities in various ways. – Bowley.

From the above definition we can know index numbers which only measures the relative change is magnitude. Index numbers relate a variable or variables to the same variable or variables pertaining to another period. In this comparison of variable the time period which is used as a basis for comparison is called as 'base period' and the other period is called as 'current period'. The Base year is of two types. They are

1. Fixed Base Year and
2. Chain Base Year

Fixed base year is a particular year which is used as a basis for comparison of a number of succeeding years. So, sometimes the base year becomes a remote period. On the other hand, the Chain base year is a varying one and is the year immediately preceding the current year.

3. 1 Characteristics of Index Numbers

1) Index numbers are specialized averages helps in comparing the changes in variables which are in different units.
2) Index numbers are expressed in percentages to show the extent of change without using percentage sign (%).
3) Relative variations are measured with the help of index numbers.
4) Index numbers are for comparison they compare changes taking place over time or between places and like categories.

Space for Hint

3.2 Uses of Index Numbers

1) The index numbers are used as end results in many situations. For example index numbers are often cited in new reports and acting as general indicators of economic condition.
2) Index numbers are used as part of an intermediate computation to understand other information latter.
3) Sales indices were used to modify and improve the estimates of future.
4) Consumer price index is used to determine the "real" buying power of money.

3.3 Types of Index Numbers

There are various types of index numbers. The most important types are as follows,

1. Price Index (*plural*: "price indices" or "price indexes") is a normalized average (typically a weighted average) of prices for a given class of goods or services in a given region, during a given interval of time. It is a statistic designed to help to compare how these prices, taken as a whole, differ between time periods or geographical locations.

   Price indices have several potential uses. For particularly broad indices, the index can be said to measure the economy's price level or a cost of living. More narrow price indices can help producers with business plans and pricing. Sometimes, they can be useful in helping to guide investment.
2. Quantity Index numbers study the changes in the volume of goods produced or consumed; for instance industrial production, agricultural production, import, export, etc. they are useful and helpful to study the output in an economy.

Space for Hint

3. Value Index numbers compare the total value of a certain period with the total value of the base period. Here the total value is equal to the price of each, multiplied by the quantity; for instance, indices of profits, sales, inventories, etc.

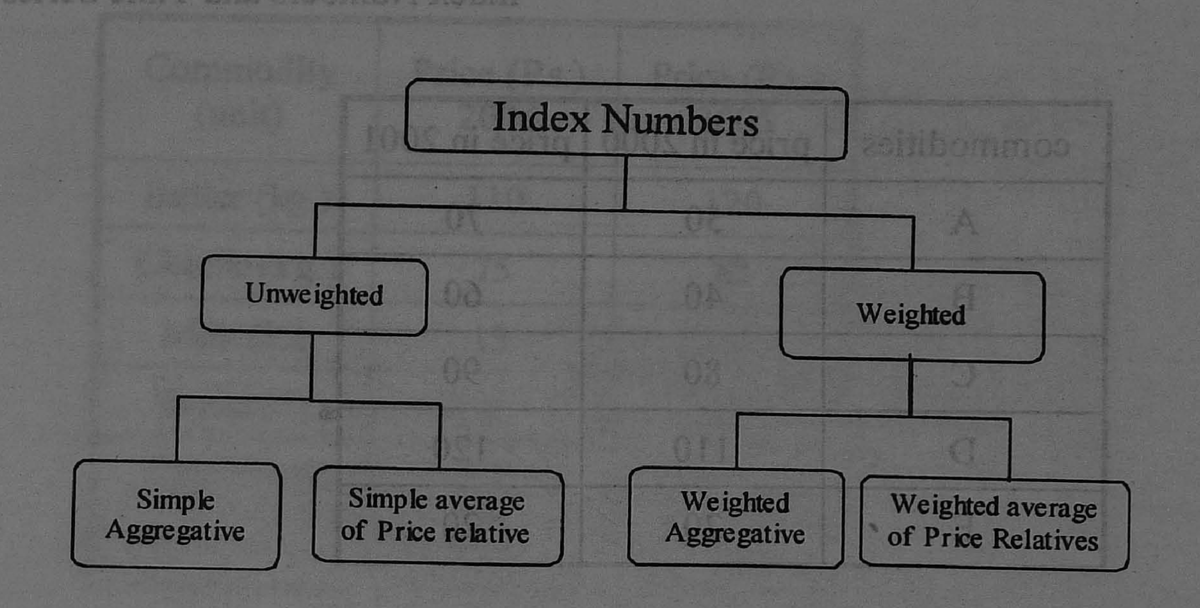
3.4 Problems Related to Index Numbers

1) Finding suitable data is difficult: Ex. If sales of a company are reported only on annual basis, it would be unable to determine the seasonal sale pattern.
2) Incomparability of indices which occurs when attempts are made to compare one index with another.
3) Inappropriate weighting of factors: In developing consumer price index, proper attention should be paid to the changes in some variables which are more important than others. This is practically very difficult and lead to it in appropriate weighting.
4) Selection of improper base: Selecting the base year may be affected by individual interest or by routine method base selection. Some times high sales may occur due to some reasons if it is selected as base. Then it will be a wrong representation. So proper care should be taken to select the base.

3.5 Construction of Index Numbers

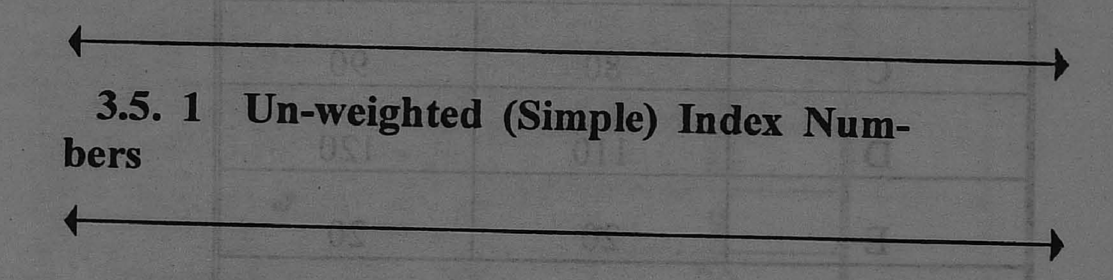
The various methods of construction of index numbers are given below:

Space for Hint



Notations : In this chapter we use the following notations.

1. p_0 = price of the commodity in the base year.
2. p_1 = price of the commodity in the current year
3. q_0 = quantity of the commodity in the base year
4. q_1 = quantity of the commodity in the current year



3.5. 1 Un-weighted (Simple) Index Numbers

1. ***Simple Aggregate Method:*** This is the simplest method of constructing the index numbers. The prices of the different commodities of the current year are added and the total is divided by the sum of the prices of the base year commodity and multiplied by 100.

Symbolically $P_{01} = \frac{\sum p_1}{\sum p_1} \times 100$

Example 3. 1

From the following data construct the aggregate index number for 2001 taking 2000 as the base.

Space for Hint

commodities	price in 2000	price in 2001
A	50	70
B	40	60
C	80	90
D	110	120
E	20	20

Solution :

We know that aggregate index number = $P_{01} = \frac{\sum p_1}{\sum p_1} \times 100$

commodities	price in 2000	price in 2001
A	50	70
B	40	60
C	80	90
D	110	120
E	20	20
Total	300	360

Now $P_{01} = \frac{\sum p_1}{\sum p_1} \times 100$

$= \frac{360}{300} \times 100$

$= 120$

Example 3. 2

From the following data construct an index number for 2005 taking 2004 as the base.

Commodity (unit)	Price (Rs.) 2004	Price (Rs.) 2005
Butter (kg.)	110	120
Cheese (kg.)	75	80
Milk (lt.)	13	13
Bread (1)	9	9
Eggs (doz)	18	20
Ghee (1 tin)	850	860

Solution :

We know that aggregate index number = $P_{01} = \frac{\sum p_1}{\sum p_1} \times 100$

Commodity (unit)	Price (rs.) 2004	Price (rs.) 2005
Butter (kg.)	110	120
Cheese (kg.)	75	80
Milk (lt.)	13	13
Bread (1)	9	9
Eggs (doz)	18	20
Ghee (1 tin)	850	860
Total	1075	1102

Now $P_{01} = \frac{\sum p_1}{\sum p_1} \times 100$

$= \frac{1102}{1075} \times 100$

$= 102.51$

Check Your Progress

(1) From the following data construct an index number for 2010 taking 2009 as the base.

Space for Hint

Space for Hint

Commodity (unit)	Price (Rs.) 2009	Price (Rs.) 2010
Rice	7	8
Wheat	3.5	3.75
Oil	40	45
Gas	78	85
flour	4.5	5.25

(Answer : $P_{01} = 110.5$)

(2) From the following data find the index number of the price relatives taking 2007 as the base year using (i) arithmetic mean and (ii) geometric mean.

Commodity (unit)	Price (Rs.) 2007	Price (Rs.) 2008
Rice	7	8
Wheat	3.5	3.75
Oil	40	45
Gas	78	85
flour	4.5	5.25

(Answer :

(i) using arithmetic mean $P_{01} = 111.92$,

(ii) using geometric mean $P_{01} = 111.87$)

(3) From the following data of the whole sale price of rice for the 5 years construct the index numbers taking (i) 1987 as the base and (ii) 1990 as the base.

Years	1987	1988	1989	1990	1991	1992
Price of Rice per kg.	5.00	6.00	6.50	7.00	7.50	8.00

Space for Hint

3. 6 Average of Price Relatives Method

The ratio of the prices $\frac{p_1}{p_2}$ is called the price relative.

The formula for finding index number for the current year

$$= p_{01}$$

$$= \frac{p_1}{p_0} \times 100$$

In the average of price relative method the average of price relatives for various items is calculated by using any one of the measures of central tendencies such as arithmetic mean, geometric mean, harmonic mean etc., Among, various central tendencies arithmetic mean and geometric mean are very common averages used in this method.

(1) The arithmetic mean index number is p_{01}

(i.e) $$p_{01} = \frac{\sum \frac{p_1}{p_0} \times 100}{n}$$

(2) The geometric mean index number is p_{01}

$$= p_{01} = \left(\prod \frac{p_1}{p_0} \right)^{1/n} \times 100$$

(i.e) $$p_{01} = \text{antilog} \left[\frac{1}{n} \sum \left(\frac{p_1}{p_0} \times 100 \right) \right]$$

Example 3. 3

From the following data compute price index by simple average method based on (i) arithmetic mean and (ii) geometric mean for 2005 taking 2004 as the base.

Space for Hint

Commodity (unit)	Price (Rs.) 2004	Price (Rs.) 2005
Butter (kg.)	110	120
Cheese (kg.)	75	80
Milk (lt.)	13	13
Bread (1)	9	9
Eggs (doz)	18	20
Ghee (1 tin)	850	860

Solution :

(i) First we shall find the price index based on arithmetic mean method.

Commodity (unit)	Price (Rs.) 2004	Price (Rs.) 2005	$\frac{p_1}{p_0}\times 100$
Butter (kg.)	110	120	109.09
Cheese (kg.)	75	80	106.67
Milk (lt.)	13	13	100.00
Bread (1)	9	9	100.00
Eggs (doz)	18	20	111.11
Ghee (1 tin)	850	860	101.18
Total	1075	1102	628.05

Thus the price index using arithmetic mean method

$$= p_{01}$$

$$= \frac{\sum \frac{p_1}{p_0}\times 100}{n}$$

$$= \frac{628.05}{6}$$

$$= 104.67$$

(ii) First we shall find the price index based on geometic mean method.

Space for Hint

Commodity (unit)	Price (Rs.) 2004	Price (Rs.) 2005	$P = \frac{p_1}{p_0} \times 100$	$\log P$
Butter (kg.)	110	120	109.09	2.0378
Cheese (kg.)	75	80	106.67	2.0280
Milk (lt.)	13	13	100.00	2.0000
Bread (1)	9	9	100.00	2.0000
Eggs (doz)	18	20	111.11	2.0458
Ghee (1 tin)	850	860	101.18	2.0051
	Total			12.1167

Thus the price index using geometric mean method

$$= p_{01}$$

$$= \text{antilog}\left[\frac{1}{n}\sum\left(\frac{p_1}{p_0} \times 100\right)\right]$$

$$= \text{antilog}\left(\frac{12.1167}{6}\right)$$

$$= 104.5785$$

$$= 104.58.$$

Example 3. 4

From the following data compute price index by simple average method based on (i) arithmetic mean, (ii) geometric mean and (iii) simple aggregative method.

Commodities	Price (Rs.) 2004	Price (Rs.) 2005
Rice	158	272
Cholam	168	326
Cambu	157	309
Ragi	155	304

Space for Hint

Solution :

(i) First we shall find the price index based on arithmetic mean method.

Commodities	Price (Rs.) 2004	Price (Rs.) 2005	$\frac{p_1}{p_0}\times 100$
Rice	158	272	172.15
Cholam	168	326	194.05
Cambu	157	309	196.82
Ragi	155	304	196.13
Total			759.1438

Thus the price index using arithmetic mean method

$$= p_{01}$$

$$= \frac{\sum \frac{p_1}{p_0}\times 100}{n}$$

$$= \frac{759.1438}{4}$$

$$= 189.79$$

(ii) First we shall find the price index based on geometric mean method.

Commodities	Price (Rs.) 2004	Price (Rs.) 2005	$P=\frac{p_1}{p_0}\times 100$	$\log P$
Rice	158	272	172.15	2.2359
Cholam	168	326	194.05	2.2879
Cambu	157	309	196.82	2.2941
Ragi	155	304	196.13	2.2925
Total				9.1104

Thus the price index using geometric mean method

$$= p_{01}$$

$$= \text{antilog}\left[\frac{1}{n}\sum\left(\frac{p_1}{p_0}\times 100\right)\right]$$

$$= \text{antilog}\left(\frac{9.1104}{4}\right)$$

$$= \text{antilog}(2.2776)$$

$$= 189.50$$

(iii) First we shall find the price index based on geometic mean method.

Commodities	Price (Rs.) 2004	Price (Rs.) 2005
Rice	158	272
Cholam	168	326
Cambu	157	309
Ragi	155	304
Total	638	1211

Thus the price index using geometric mean method

$$= p_{01}$$

$$= \frac{\sum p_1}{\sum p_0}\times 100$$

$$= \frac{1211}{638}$$

$$= 189.81$$

Example 3. 5

For the data given below calculate the index numbers taking (i) 2001 as the base year and (ii) 2005 as base year.

Space for Hint

Space for Hint

years	2001	2002	2003	2004	2005	2006	2007	2008	2009
price of wheat per kg.	4	5	6	7	8	10	9	10	11

Solution :

Construction of index numbers taking 2001 as base year

Years	Price of wheat per kg.	Index Numbers
2001	4	100
2002	5	$\frac{4}{5}\times100=125$
2003	6	$\frac{6}{5}\times100=150$
2004	7	$\frac{7}{5}\times100=175$
2005	8	$\frac{8}{5}\times100=200$
2006	10	$\frac{10}{5}\times100=250$
2007	9	$\frac{9}{5}\times100=225$
2008	10	$\frac{10}{5}\times100=250$
2009	11	$\frac{11}{5}\times100=275$

Construction of index numbers taking 2001 as base year

Years	Price of wheat per kg.	Index Numbers
2001	4	$\frac{4}{8}\times 100 = 50$
2002	5	$\frac{4}{8}\times 100 = 62.50$
2003	6	$\frac{6}{8}\times 100 = 75$
2004	7	$\frac{7}{8}\times 100 = 87.50$
2005	8	$\frac{8}{8}\times 100 = 100$
2006	10	$\frac{10}{8}\times 100 = 125$
2007	9	$\frac{9}{8}\times 100 = 112.50$
2008	10	$\frac{10}{8}\times 100 = 125$
2009	11	$\frac{11}{8}\times 100 = 137.50$

Example 3. 6

Construct the whole sale price index number for 2001 and 2002 from the following data by considering 2000 as the base year.

Commodity	Whole Sale Price In Rs. per quintal		
	2000	2001	2002
Rice	700	750	825
Wheat	540	575	600
Ragi	300	325	310
Cholam	250	280	295
Flour	320	330	335
Ravai	325	350	360

Space for Hint

Space for Hint

Solution :

Construction of index numbers taking 2001 as base year

Commodity	Whole Sale Price in Rs. per quintal			relative for 2001	relative for 2002
	2000 p_0	2001 p_1	2002 p_1		
Rice	700	750	825	$\frac{750}{700}\times100$ = 107.14	$\frac{825}{700}\times100$ = 117.86
Wheat	540	575	600	$\frac{575}{540}\times100$ = 106.48	$\frac{600}{540}\times100$ =111.11
Ragi	300	325	310	$\frac{325}{300}\times100$ = 108.33	$\frac{310}{300}\times100$ = 103.33
Cholam	250	280	295	$\frac{280}{250}\times100$ = 112.00	$\frac{295}{250}\times100$ = 118.00
Flour	320	330	335	$\frac{330}{320}\times100$ = 103.13	$\frac{325}{325}\times100$ = 104.69
Ravai	325	350	360	$\frac{350}{325}\times100$ = 107.69	$\frac{360}{325}\times100$ =110.77
Total				644.77	665.76
Index number using A.M.				107.46	110.96

Thus the index number for 2001 as base year 2000 = 107.46

and the index number for 2002 as base year 2000 = 110.96

Example 3. 7

From the following data find (i) fixed base index numbers with 2003 as the base year and (ii) chain base index numbers.

Space for Hint

Commodity	Price in Rs.				
	2001	2002	2003	2004	2005
I	2	3	5	7	6
II	8	10	12	4	18
III	4	3	7	9	12

Solution :

Construction of fixed base index numbers taking 2001 as base year

Commodity	Price in Rs.				
	2001	2002	2003	2004	2005
I	100	150	250	350	300
II	100	125	150	50	225
III	100	75	175	225	300
Total	300	350	575	625	825
Index number (A.M.)	100	116.67	191.67	208.33	275

Workings :

Fixed base index number for the years 2002, 2003, 2004 and 2005 based on 2001 for the commodity I be calculated as

$\frac{3}{2}\times100=150$, $\frac{5}{2}\times100=250$, $\frac{7}{2}\times100=350$ and $\frac{6}{2}\times100=300$.

Similarly for the years 2002, 2003, 2004 and 2005 based on 2001 for the commodity II be calculated as

$\frac{10}{8}\times100=125$, $\frac{12}{8}\times100=150$, $\frac{4}{8}\times100=50$, $\frac{18}{8}\times100=225$.

Space for Hint

Construction of chain base index numbers.

Commodity	Price in Rs.				
	2001	2002	2003	2004	2005
I	100	$\frac{3}{2}\times 100$ = 150	$\frac{5}{3}\times 100$ = 166.67	$\frac{7}{5}\times 100$ = 140	$\frac{6}{7}\times 100$ = 85.71
II	100	$\frac{10}{8}\times 100$ = 125	$\frac{12}{10}\times 100$ = 120	$\frac{4}{12}\times 100$ = 33.33	$\frac{18}{4}\times 100$ = 450
III	100	$\frac{3}{4}\times 100$ = 75	$\frac{7}{3}\times 100$ = 233.33	$\frac{9}{7}\times 100$ = 128.57	$\frac{12}{9}\times 100$ = 133.33
Total	300	350	520	301.90	669.05
Index number (A.M.)	100	116.67	173.33	100.63	223.02

3. 7 Weighted Index Number

The unweighted index numbers studied in the earlier sections are not unweighted in the true sense of the term. Actually we assign equal importance to all the items included in the index and as such they are in reality weighted, weights being implicit rather than explicit.

Weighted index numbers are of two types. The purpose of weighting is to make the index numbers more perspective and to give more importance to them. They are (i) Weighted Aggregate Index Numbers and (ii) Weighted Average of Price Relatives.

Space for Hint

3.7. 1 Weighted Aggregative Index Number

Weighted aggregative index numbers are of the simple aggregative type with the fundamental difference that weights are assigned to the various items included in the index. There are various methods of assigning weights and consequently a large number of formulae for constructing index numbers have been devised. Some of the important index numbers are

1. Laspeyre's index number
2. Paasche's index number
3. Dorbish and Bowley's index number
4. Fisher's index number
5. Marshall – Edgeworth index number
6. Kelly's index number

Laspeyre's index number :

In this method the price of the commodities in the base year as well as the current year are known and the they are weighted by the quantity used in the base year.

(i.e) $$P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

Paasche's index number :

In this method the price of the commodities in the base year as well as the current year are known and the they are weighted by the quantity used in the current year.

(i.e) $$P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

Space for Hint

Dorbish and Bowley's method:

This is an index number got by the arithmetic mean of Laspeyre's and Paasche's index numbers.

(i.e) $$P_{01} = \frac{1}{2}\left[\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}\right] \times 100$$

Fisher's Ideal method:

Fisher's price index number is given by the geometric mean of Laspeyre's and Paasche's index numbers.

(i.e) $$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

Marshall – Edgeworth method:

In Marshall – Edgeworth's index number, the arithmetic mean of base year and current year quantities are taken as the weights.

(i.e) $$P_{01} = \frac{\Sigma p_1(q_0 + q_1)}{\Sigma p_0(q_0 + q_1)} \times 100$$

Kelly's method:

Kelly's index number uses quantities of some period (which is neither the base year nor the current year) as weights. This weight is kept constant for all periods.

(i.e) $$P_{01} = \frac{\Sigma p_1 q}{\Sigma p_0 q} \times 100$$ where $q = \frac{q_0 + q_1}{2}$

Example 3. 8

Construct the index numbers of prices from the following data using

(i) Laspeyre's index number

(ii) Paache's index number

(iii) Bowley's index number

(iv) Fisher's index number

(v) Marshall-Edgeworth index number

Commodity	Base year		Current year	
	Price	Quantity	Price	Quantity
A	2	8	4	6
B	5	10	6	5
C	4	14	5	10
D	2	19	2	13

Solution :

Commod-ity	Base year		Current year		p_0q_0	p_1q_0	p_0q_1	p_1q_1
	Price p_0	Quantity q_0	Price p_1	Quantity q_1				
A	2	8	4	6	16	32	12	24
B	5	10	6	5	50	60	25	30
C	4	14	5	10	56	70	40	50
D	2	19	2	13	38	38	26	26
				Total	160	200	103	130

Now Laspeyre's index number

$= P_{01}$

$= \dfrac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$

$= \dfrac{200}{160} \times 100$

$= 125$

Space for Hint

Space for Hint

and Paache's index number

$$= P_{01}$$

$$= \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

$$= \frac{130}{103} \times 100$$

$$= 126.21$$

and Bowley's index number

$$= P_{01}$$

$$= \frac{1}{2}\left[\frac{\sum p_1 q_0}{\sum p_0 q_0} + \frac{\sum p_1 q_1}{\sum p_0 q_1}\right] \times 100$$

$$= \frac{1}{2}\left[\frac{200}{160} + \frac{130}{103}\right] \times 100$$

$$= 125.61$$

and Fisher's index number

$$= P_{01}$$

$$= \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

$$= \sqrt{\frac{200}{160} \times \frac{130}{103}} \times 100$$

$$= 125.61$$

and Marshall-Edgeworth's index number

$$= P_{01}$$

$$= \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \times 100$$

$$= \frac{1}{2}\left[\frac{200}{160} + \frac{130}{103}\right] \times 100$$

$$= 125.48$$

Example 3. 9

For the given data find the different weighted index numbers.

Commodity	Base year		Current year	
	Price	Quantity	Price	Quantity
A	6	50	10	56
B	2	100	2	120
C	4	60	6	60
D	10	30	12	24
E	8	40	12	26

Solution :

Commodity	Base year		Current year		p_0q_0	p_1q_0	p_0q_1	p_1q_1
	Price p_0	Quantity q_0	Price p_1	Quantity q_1				
A	6	50	10	56	300	500	336	560
B	2	100	2	120	200	200	240	240
C	4	60	6	60	240	360	240	360
D	10	30	12	24	300	360	240	288
E	8	40	12	26	320	480	208	312
				Total	1360	1900	1264	1760

Space for Hint

Now Laspeyre's index number

$$= P_{01}$$

$$= \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

$$= \frac{1900}{1360} \times 100$$

$$= 139.71$$

and Paache's index number

$$= P_{01}$$

$$= \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

$$= \frac{1760}{1264} \times 100$$

$$= 139.24$$

and Bowley's index number

$$= P_{01}$$

$$= \frac{1}{2}\left[\frac{\sum p_1 q_0}{\sum p_0 q_0} + \frac{\sum p_1 q_1}{\sum p_0 q_1}\right] \times 100$$

$$= \frac{1}{2}\left[\frac{1900}{1360} + \frac{1760}{1264}\right] \times 100$$

$$= 139.47$$

and Fisher's index number

$$= P_{01}$$

$$= \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

$$= \sqrt{\frac{1900}{1360} \times \frac{1760}{1264}} \times 100$$

$$= 139.47$$

and Marshall-Edgeworth's index number

$= P_{01}$

$$= \frac{\sum p_1q_0 + \sum p_1q_1}{\sum p_0q_0 + \sum p_0q_1} \times 100$$

$$= \left[\frac{1900}{1360} + \frac{1760}{1264}\right] \times 100$$

$= 139.48$

Example 3. 10

For the given data find the different weighted index numbers.

Commodity	Base year		Current year	
	Price	Value	Price	Value
A	10	100	12	144
B	15	75	20	120
C	8	80	10	110
D	20	60	25	50
E	50	500	60	540

Solution :

Commodity	Base year		Current year		p_0q_0	p_1q_0	p_0q_1	p_1q_1
	Price p_0	Quantity q_0	Price p_1	Quantity q_1				
A	10	10	12	12	100	120	120	144
B	15	5	20	6	75	100	90	120
C	8	10	10	11	80	100	88	110
D	20	3	25	2	60	75	40	50
E	50	10	60	9	500	600	450	540
				Total	815	995	788	964

Space for Hint

Space for Hint

Now Laspeyre's index number

$$= P_{01}$$

$$= \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

$$= \frac{788}{815} \times 100$$

$$= 122.09$$

and Paache's index number

$$= P_{01}$$

$$= \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

$$= \frac{964}{995} \times 100$$

$$= 122.34$$

and Bowley's index number

$$= P_{01}$$

$$= \frac{1}{2}\left[\frac{\sum p_1 q_0}{\sum p_0 q_0} + \frac{\sum p_1 q_1}{\sum p_0 q_1}\right] \times 100$$

$$= \frac{1}{2}\left[\frac{788}{815} + \frac{964}{995}\right] \times 100$$

$$= 122.21$$

and Fisher's index number

$$= P_{01}$$

$$= \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

$$= \sqrt{\frac{788}{815} \times \frac{964}{995}} \times 100$$

$$= 122.21$$

and Marshall-Edgeworth's index number

$= P_{01}$

$$= \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \times 100$$

$$= \left[\frac{788}{815} + \frac{964}{995}\right] \times 100$$

$= 122.21$

Example 3. 11

Find the value of x in the following data if the ratio between Laspyre's and Paache's index numbers is $28:27$

Commodity	Base year		Current year	
	Price	Value	Price	Value
A	1	10	2	5
B	1	5	x	2

Solution :

Commodity	Base year		Current year		$p_0 q_0$	$p_1 q_0$	$p_0 q_1$	$p_1 q_1$
	Price p_0	Quantity q_0	Price p_1	Quantity q_1				
A	1	10	2	5	10	20	5	10
B	1	5	x	2	5	$5x$	2	$2x$
				Total	15	$20+5x$	7	$10+2x$

Laspeyre's index number $= L_{01}$

$$= \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

$$= \frac{20+5x}{15} \times 100$$

Space for Hint

Space for Hint

Paache's index number = P_{01}

$$= \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

$$= \frac{10+2x}{7} \times 100$$

Given that $L_{01} : P_{01} = 28:27$

(i.e) $\frac{20+5x}{15_5} \times \frac{7}{10+2x} = \frac{28_4}{27_9}$

(i.e) $180+45x = 200+40x$

(i.e) $x = 4$

Thus the required value is 4.

Check Your Progress

(1) For the given data find the different weighted index numbers.

Commodity	Base year		Current year	
	Price	Quantity	Price	Quantity
A	10	25	12	30
B	8	21	9	25
C	4.5	28	6.5	35
D	3.5	16	4	20

(2) For the given data find the different weighted index numbers.

Commodity	Base year		Current year	
	Price	Quantity	Price	Quantity
A	2	10	3	30
B	5	16	6.5	11
C	3.5	18	4	16
D	7	21	9	25
E	3	11	3.5	20

(3) For the given data find the different weighted index numbers.

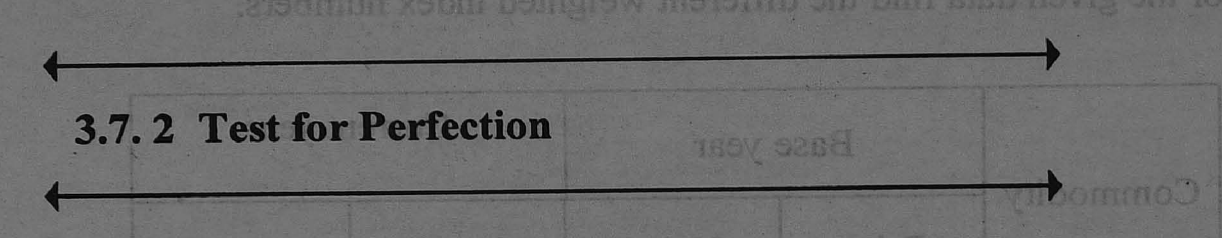
Commodity	Base year		Current year	
	Price	Value	Price	Value
A	8	200	65	1950
B	20	1400	30	1650
C	5	80	20	900
D	10	360	15	300
E	27	2160	10	600

(4) For the given data find the Fisher's index numbers.

Commodity	Base year		Current year	
	Price	Quantity	Price	Quantity
A	5	14	3	8
B	8	18	6	25
C	3	25	1	40
D	15	36	12	48
E	9	14	7	18
F	7	13	5	19

Space for Hint

Space for Hint



3.7. 2 Test for Perfection

Several formulae have been suggested for constructing index numbers and the problem is that of selecting the most appropriate one in a given situation. The following test are suggested for choosing an appropriate index number :

(1) Time reversal test, (2) Factor reversal test.

Time Reversal Test :

Reversibility is an important property that an index number should posses. A good index number should satisfy the time reversal tests. In the words of Irving Fisher, " The formula for calculating an index number should be such that it gives the same ratio between one point of comparison and the other , no matter which of the two is taken as the base : or putting it in another way , the index number reckoned forward should be reciprocal of the one reckoned backward." One of the advantages claimed in favor of Fisher's formula is that it makes the index number reversible.

The time reversal test shows that the following equations hold good

$$\boxed{P_{01} \times P_{10} = 1}$$

where P_{01} is the index for time "1" on time "0" as base and P_{10} is the index for time "0" on time "1" as base.

If the product is not unity, then we can say there is a time bias in the method.

Example 3. 12

Verify that Laspeyre's index number satisfies time reversal test.

Solution :

We know that Laspeyre's index number

$$= P_{01}$$

$$= \frac{\sum p_1 q_0}{\sum p_0 q_0}$$

Space for Hint

and $P_{10} = \frac{\sum p_0 q_0}{\sum p_1 q_1}$

Now $P_{01} \times P_{10} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_0 q_0}{\sum p_1 q_1}$

$\neq 1.$

Therefore Laspeyre's index number does not satisfy time reversal test.

Example 3. 13

Verify that Paache's index number satisfies time reversal test.

Solution :

We know that Paache's index number

$= P_{01}$

$= \frac{\sum p_1 q_1}{\sum p_0 q_1}$

and $P_{10} = \frac{\sum p_0 q_0}{\sum p_1 q_0}$

Now $P_{01} \times P_{10} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}$

$\neq 1.$

Therefore Paache's index number does not satisfy time reversal test.

Example 3. 14

Verify that Fisher's index number satisfies time reversal test.

Solution :

We know that Fisher's index number

$= P_{01}$

$= \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}}$

and $P_{10} = \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}}$

Now $P_{01} \times P_{10} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}}$

$= 1.$

Therefore Fisher's index number satisfies time reversal test.

Space for Hint

2. Factor Reversal Test :

Another basic test is that the formula for index number ought to permit to interchanging the prices and quantities without giving inconsistent results i.e., the two results multiplied together should give the true value ratio. A good index number should satisfy not only the time reversal test, but also the factor reversal test. A good index number should allow time reversibility, interchange of the base year and the current year, without giving inconsistent results.

(i.e) $$P_{01} \times Q_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

where P_{01} is the index for time "1" on time "0" as base and Q_{01} is the quantity index for time "1" on time "0" as base.

If the product is not unity, then we can say there is a time bias in the method.

Example 3. 15

Verify that Fisher's index number satisfies factor reversal test.

Solution :

We know that Fisher's index number

$$= P_{01}$$

$$= \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$$

$$\text{and } Q_{01} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}}$$

$$\text{Now } P_{01} \times Q_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}}$$

$$= \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}.$$

Therefore Fisher's index number satisfies time reversal test.

Note : ***An index number satisfies both time and factor reversal tests then that index number is called ideal index number.***

Example 3. 16

For the given data verify that Fisher's index number is an ideal index number.

Commodity	Base year		Current year	
	Price	Value	Price	Value
A	10	100	12	144
B	15	75	20	120
C	8	80	10	110
D	20	60	25	50
E	50	500	60	540

Solution :

Commodity	Base year		Current year		p_0q_0	p_1q_0	p_0q_1	p_1q_1
	Price p_0	Quantity q_0	Price p_1	Quantity q_1				
A	10	10	12	12	100	120	120	144
B	15	5	20	6	75	100	90	120
C	8	10	10	11	80	100	88	110
D	20	3	25	2	60	75	40	50
E	50	10	60	9	500	600	450	540
				Total	815	995	788	964

Space for Hint

Space for Hint

First we shall prove that Fisher's index number satisfies time reversal test.

Now the Fisher's index number, $P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$

$= \sqrt{\frac{995}{815} \times \frac{964}{788}}$

$= 1.22$

and $P_{10} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}}$

$= \sqrt{\frac{788}{964} \times \frac{815}{995}}$

$= 0.82$

Now $P_{01} \times P_{10} = 1.22 \times 0.82$

$= 1$

Hence Fisher's index number satisfies time reversal test.

First we shall prove that Fisher's index number satisfies factor reversal test.

Now the Fisher's index number, $P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$

$= \sqrt{\frac{995}{815} \times \frac{964}{788}}$

$= 1.22$

and $P_{10} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}}$

$= \sqrt{\frac{788}{815} \times \frac{964}{995}}$

$= 0.97$

Now $P_{01} \times Q_{01} = 1.22 \times 0.97$

$= 1.18$

and $\frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} = \frac{964}{815}$

$= 1.18$

Thus $P_{01} \times Q_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$.

Hence Fisher's index number satisfies factor reversal test.

Thus Fisher's index number is an ideal index number.

Example 3. 17

For the given data verify that Fisher's index number is an ideal index number.

Commodity	Base year		Current year	
	Price	Value	Price	Value
A	6	50	10	56
B	2	100	2	120
C	4	60	6	60
D	10	30	12	24
E	8	40	12	26

Solution :

Commodity	Base year		Current year		p_0q_0	p_1q_0	p_0q_1	p_1q_1
	Price p_0	Quantity q_0	Price p_1	Quantity q_1				
A	6	50	10	56	300	500	336	560
B	2	100	2	120	200	200	240	240
C	4	60	6	60	240	360	240	360
D	10	30	12	24	300	360	240	288
E	8	40	12	26	320	480	208	312
				total	1360	1900	1264	1760

Space for Hint

Space for Hint

First we shall prove that Fisher's index number satisfies time reversal test.

Now the Fisher's index number, $P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$

$= \sqrt{\frac{1900}{1360} \times \frac{1760}{1264}}$

$= 1.39$

and $P_{10} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}}$

$= \sqrt{\frac{1264}{1760} \times \frac{1360}{1900}}$

$= 0.72$

Now $P_{01} \times P_{10} = 1.39 \times 0.72$

$= 1$

Hence Fisher's index number satisfies time reversal test.

First we shall prove that Fisher's index number satisfies factor reversal test.

Now the Fisher's index number, $P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$

$= \sqrt{\frac{1900}{1360} \times \frac{1760}{1264}}$

$= 1.39$

and $P_{10} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}}$

$= \sqrt{\frac{1264}{1360} \times \frac{1760}{1264}}$

$= 1.14$

Now $P_{01} \times Q_{01} = 1.39 \times 1.14$

$= 1.59$

and $\frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} = \frac{1760}{1360}$

$= 1.59$

Thus $P_{01} \times Q_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$.

Hence Fisher's index number satisfies factor reversal test.

Thus Fisher's index number is an ideal index number.

Check Your Progress

Verify the Fisher's index number is an ideal index number for the following data.

Commodity	Base year		Current year	
	Price	Value	Price	Value
A	8	60	12	66
B	4	120	4	160
C	6	60	8	90
D	15	45	14	52
E	10	50	14	36

3.7. 3 Weighted Average of Price Relative Method

In the weighted aggregative methods discussed in the earlier sections price relatives were not computed. However, like unweighted relatives method it is possible to compute weighted average of relatives. For purposes of averaging we use either arithmetic mean or the geometric mean.

The formula for weighted average of relative index number using arithmetic mean is $P_{01} = \frac{\sum PV}{\sum V}$ where P = price relative and V = value weights.

Note : In the above formula P can be calculated from $\frac{p_1}{p_0} \times 100$

(i.e) $P = \frac{p_1}{p_0} \times 100$ and $V = p_0 q_0$.

The formula for weighted average of relative index number using geoemtric mean is $P_{01} = anti\log\left\{\frac{\sum V \log P}{\sum V}\right\}$ where P = price relative and V = value weights.

Space for Hint

Space for Hint

Note : In the above formula P can be calculated from $\frac{p_1}{p_0}\times 100$

(i.e) $P=\frac{p_1}{p_0}\times 100$ and $V=p_0q_0$.

Example 3. 18

From the following data compute price index by applying weighted average of price relative method using (i) arithmetic mean and (ii) geometric mean.

Commodity	p_0	q_0	p_1
Sugar	28	20	38
Flour	22	40	34
Milk	30	10	36

Solution :

Index number using weighted arithmetic mean of price relatives

Commodity	p_0	q_0	p_1	$V=p_0q_0$	$P=\frac{p_1}{p_0}\times 100$	PV
Sugar	28	20	38	560	111.11	62222
Flour	22	40	34	880	116.67	102667
Milk	30	10	36	300	106.67	32000
	Total			1740		196889

Now $P_{01} = \frac{\Sigma PV}{\Sigma V}$

$= \frac{196889}{990}$

$= 113.15$

Index number using weighted geometric mean of price relatives

Commodity	p_0	q_0	p_1	$V = p_0 q_0$	$P = \frac{p_1}{p_0} \times 100$	$V \log P$	PV
Sugar	28	20	38	560	111.11	2.0458	1146
Flour	22	40	34	880	116.67	2.0669	1819
Milk	30	10	36	300	106.67	2.0280	608
		Total		1740			3573

Now $P_{01} = anti\log\left\{\frac{\sum V \log P}{\sum V}\right\}$

$= anti\log\left\{\frac{3573}{990}\right\}$

$= 113.09$.

Check Your Progress

From the following data compute price index by applying weighted average of price relative method using (i) arithmetic mean and (ii) geometric mean.

Commodity	p_0	q_0	p_1
Sugar	18	20	28
Flour	12	40	24
Milk	20	10	26

Space for Hint

3. 8 Consumer Price Index Number

Consumer index number are useful for wage negotiations and wage contracts. Dearness allowance is calculated based on the cost of living index numbers. Even for family budgets cost of living index numbers are calculated taking into consideration the importance of the consumption of articles. Generally items on which the money is spent are classified into certain heads such as Food, clothing, fuel, rent etc., each of these groups can further subdivided. Suitable weights can be associated as per the relative importance of the items.

Formula to find the cost of living index numbers based on

(i) Aggregate expenditure method

Cost of living index number = I_{01}

(i.e) $$I_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

(ii) Family budget method

Cost of living index number = I_{01}

(i.e) $$I_{01} = \frac{\sum PV}{\sum V}$$

where P = price relative $= \frac{p_1}{p_0} \times 100$

and V = value weights = $p_0 q_0$

Also $$I_{01} = \frac{\sum PW}{\sum W}$$

where P = price relative $= \frac{p_1}{p_0} \times 100$

and W = value weights = $p_0 q_0$

Space for Hint

Example 3. 19

Calculate the index number of prices for 2006 on the basis of 2004 from the data given below.

commodity	weights	Price per unit 2004	price per unit 2006
A	40	80	85
B	25	60	55
C	5	345	50
D	20	35	40
E	10	25	20

Solution :

We know that consumer price index number $= \dfrac{\Sigma PW}{\Sigma W}$

Commodity	Weights W	Price per unit 2004	Price per unit 2006	$P = \dfrac{p_1}{p_0} \times 100$	PW
A	40	80	85	106.25	4250.00
B	25	60	55	91.67	2291.67
C	5	34	50	147.06	735.29
D	20	35	40	114.29	2285.71
E	10	25	20	80.00	800.00
Total	100			539.26	10362.68

Now consumer price index number $= \dfrac{\Sigma PW}{\Sigma W}$

$$= \frac{10362.68}{539.26}$$

$$= 103.63$$

Space for Hint

Example 3. 20

An enquiry into the budgets of the middle class families in a city in India gave the following information.

	Food	Rent	Clothing	Fuel	Misc
Weights	35%	15%	20%	10%	20%
Prices in 2001	1800	500	600	100	500
Prices in 2002	2000	700	900	130	550

What changes in cost of living index of 2002 as compared with that of 2001 are seen?

Solution :

We know that cost of living index number $= \dfrac{\Sigma PW}{\Sigma W}$

Now

	Weights	Prices in 2001	Prices in 2002	P	W	PW
Food	35%	1800	2000	111.11	35	3888.89
Rent	15%	500	700	140.00	15	2100.00
Clothing	20%	600	900	150.00	20	3000.00
Fuel	10%	100	130	130.00	10	1300.00
Misc	20%	500	550	110.00	20	2200.00
Total					100	12488.89

Thus the cost of living index number $= \dfrac{\Sigma PW}{\Sigma W}$

$$= \frac{12488.89}{100.00}$$

$$= 124.90$$

Thus the prices in 2002 compared with the price in 2001 has risen to 24.9%.

Check Your Progress

An enquiry into the budgets of the middle class families in a city in India gave the following information.

	Food	Rent	Clothing	Fuel	Misc
Weights	40%	20%	20%	10%	10%
Prices in 2008	2800	1500	1600	1100	1500
Prices in 2009	3000	1700	1900	1130	1550

What changes in cost of living index of 2009 as compared with that of 2008 are seen?

Space for Hint

3.8.1 CONVERSION OF CHAIN BASE INDEX NUMBER INTO FIXED BASE INDEX NUMBER AND CONVERSELY

The fixed base index number shows the changes in the level of phenomenon as compared to a fixed year as base year. But the chain base index number compares the level from the preceding year. In some situations it is necessary that to convert the chain base index number into a fixed base index number directly without referring to the actual prices.

Thus fixed base index of the current year

$$\text{Fixed base index of the current year} = \frac{\left(\text{chain base index of the current year}\right) \times \left(\text{fixed base index of the preceding year}\right)}{100}$$

and

$$\text{Chain base index number of the current year} = \frac{\left(\text{fixed base index of the current year}\right) \times 100}{\text{fixed base index of the previous year}}$$

Space for Hint

Example 3. 21

Convert the following chain base index numbers into fixed base index numbers.

Years	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Chain index	100	112.8	86.4	102.6	120.5	105.3	103.3	109.8	88.4	75.8

Solution :

We know that

$$\text{Fixed base index of the current year} = \frac{\left(\begin{array}{l}\text{chain base index of}\\ \text{the current year}\end{array}\right)\times\left(\begin{array}{l}\text{fixed base index of}\\ \text{the preceding year}\end{array}\right)}{100}$$

Now the third column of the following table show the fixed base index number

Years	Chain index	Fixed base index number
1993	100	100
1994	112.8	$\frac{112.8\times100}{100}=112.8$
1995	86.4	$\frac{112.8\times86.4}{100}=97.5$
1996	102.6	100.0
1997	120.5	120.5
1998	105.3	126.9
1999	103.3	131.1
2000	109.8	143.9
2001	88.4	127.2
2002	75.8	96.4

Example 3. 22

Convert the following chain base index numbers into fixed base index numbers.

Years	1996	1997	1998	1999	2000
Chain index	80	110	120	90	140

Solution :

We know that

$$\text{Fixed base index of the current year} = \frac{\left(\begin{array}{c}\text{chain base index of}\\ \text{the current year}\end{array}\right) \times \left(\begin{array}{c}\text{fixed base index of}\\ \text{the preceding year}\end{array}\right)}{100}$$

Now the third column of the following table show the fixed base index number

Years	Chain index	Fixed base index number
1996	80	80
1997	110	$\frac{110 \times 80}{100} = 88$
1998	120	$\frac{120 \times 88}{100} = 105.6$
1999	90	95.04
2000	140	133.06

Check Your Progress

Convert the following chain base index numbers into fixed base index numbers.

Years	2000	2001	2002	2003	2004	2005
Chain index	90	105	102	98	120	125

Space for Hint

Space for Hint

Example 3. 23

From the fixed base index numbers for the data given below prepare chain base index numbers.

Years	2000	2001	2002	2003	2004	2005
Chain index	94	98	102	95	98	100

Solution :

We know that

$$\text{Chain base index number of the current year} = \frac{\left(\text{fixed base index of the current year}\right) \times 100}{\text{fixed base index of the previous year}}$$

Now the third column of the following table shows the chain base index numbers.

Years	Fixed base index number	Chain index
2000	94	100.00
2001	98	104.26
2002	102	104.08
2003	95	93.14
2004	98	103.16
2005	100	102.04

Example 3. 24

From the fixed base index numbers for the data given below prepare chain base index numbers.

Years	2000	2001	2002	2003	2004	2005
Fixed base index number	100	112.8	97.4	100	120.5	126.1

Space for Hint

Solution :

We know that

$$\text{Chain base index number of the current year} = \frac{\left(\text{fixed base index of the current year}\right) \times 100}{\text{fixed base index of the previous year}}$$

Now the third column of the following table show the chain base index numbers.

Years	Fixed base index number	Chain index
2000	100	100.00
2001	112.8	112.80
2002	97.4	86.35
2003	100	102.67
2004	120.5	120.50
2005	126.1	104.65

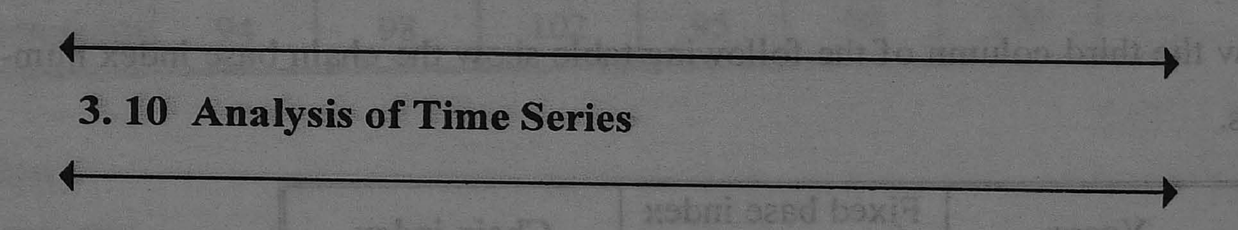
3. 9 Limitations of Index Numbers

Even though index numbers are very important in business and economic activities, they have their own limitations; they are:

1. There may be error in each stage of the construction of the index number, namely, selection of commodities, selection of base period, selection of weight, etc.
2. Index numbers may not represent the exact change in price level, because they are based on sample data.
3. Tastes, habits and customs of people change in course of time and may make the weighting not suitable for the present data.

Space for Hint

4. In each index there is an index error, because there is no formula for measuring the price change. So there is the formula error. Hence it will not be a representative one.
5. By selecting a suitable year as the base year, selfish persons may get their desired results.



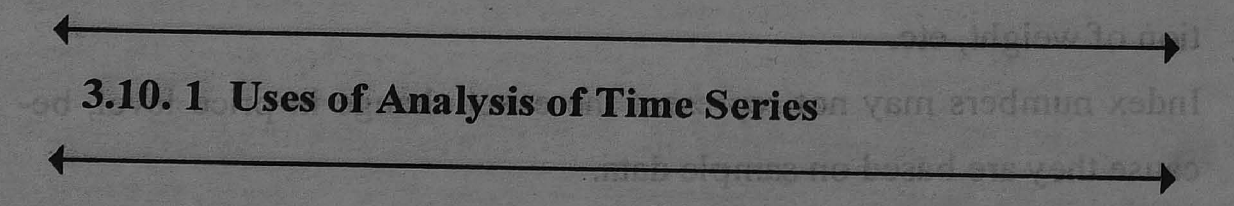
3. 10 Analysis of Time Series

Forecasting or predicting is an essential tool in any decision-making process. It uses vary from determining inventory requirements for a local shoe store to estimating the annual sales of vivo games. The quality of the forecasts management can make is strongly related to the information that can be extracted and used from past data. Time series analysis is one quantitative method we use to determine patterns in data collected over time

Time series analysis is used to detect patterns of change in statistical information over regular intervals of time. We project these patterns to arrive at an estimate for the future. Thus time series analysis helps us cope with uncertainty about the future

Definition :

An arrangement of statistical data in accordance with time of occurrence or in chronological order is called a time series. Thus when we observe numerical data at different points of time the set of observation is known as time series.



3.10. 1 Uses of Analysis of Time Series

Time series analysis is useful in different fields like economics, science, research work, etc, because of the following reasons.

1. It helps in understanding the past behavior
2. It helps in planning and forecasting.
3. Comparison between data of one period with another period is possible
4. It is useful not only to economists but also to the businessman.
5. It helps in evaluating current accomplishments

3. 11 component s of time series or types of variations

There are four basic types of variations and these are called the components or elements of time series. They are

1. Secular Trend
2. Seasonal variation
3. Cyclical variation and
4. Irregular variations

Secular Trend

The first type of variation in a time series is known as secular trend. The value of the variable tends to increase or decrease over a long period of time. The steady increase in the cost of living recorded by the consumer price index is an example of secular trend. From year individual year, the cost of living varies a great deal, but we examine a long-term period, we see that the trend is toward a steady increase It is an increasing but fluctuating time series.

Uses of Trend

1. The trend describes the basic growth tendency ignoring short-term fluctuations.
2. It describes the pattern of behavior, which characterized the series in the past.
3. Future behaviour can be forecasted in the assumption that past behaviour will continue in the future also.
4. Trend analyses facilitate us to compare two or more time series over different period of time and this helps to draw conclusions about them.

Space for Hint

Space for Hint

2. Seasonal Variation

The third kind of change in time series data is seasonal variation. As we might expect from the name, seasonal variation involves patterns of change with in a year that tend to be repeated from year to year. For example, physician can expect a substantial increase in the number of flu cases every winter and of poison ivy every summer. Because these are regular patterns, they are useful in forecasting the future. The seasonal variation may occur due to

Climate and natural forces.

The result of natural forces like climate is causing seasonal variation. Umbrellas are sold more in rainy season. In winter season, sale of the woolen clothes will increase. In hot season, the sales of ice, ice-cream, fruit salad etc. will increase. Thus climate and weather play an important role in seasonal movement. Agricultural production depends upon the monsoon.

Customs and habits.

Man-made conventions are the customs, habits, fashion, etc. There is the custom of wearing new clothes, preparing sweets and buying crackers for Deepavali, Onam, Christmas, etc. At that time, there is more demand for cloth, sweets and crackers. It will happen every year. In marriage season, the price of gold will increase. Seasonal variations are useful to businessmen, agriculturist, sales managers and producers.

3. Cyclical Fluctuation

The second type of variation seen in a time series is cyclical fluctuation the most common example of cyclical fluctuation is the business cycle. Over time, there are years when the business cycle hits a peak about the trend line. At other times, business activity is likely to slump hitting a low point below the trend line. The time between hitting peaks or falling to low points is at least one year, and it can be as many as 15 to 20 years. Example cyclical movements do not follow any regular pattern but move in a somewhat unpredictable manner.

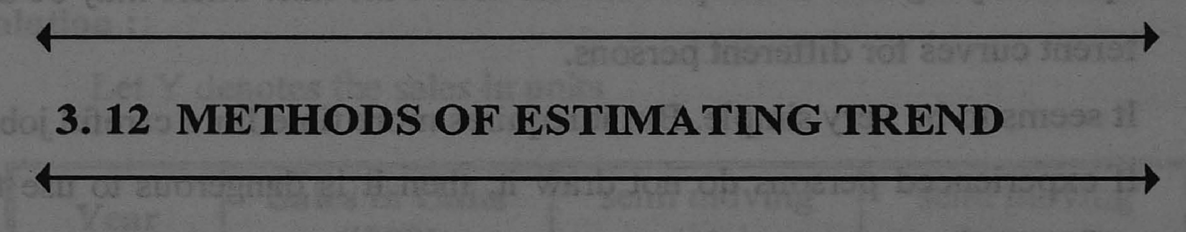
Space for Hint

The purpose of studying the cyclical variation is:

1. One can easily study the character of business fluctuations. Good policies can be formulated at stabilizing the level of business activity.
2. Businessman can take timely steps in maintaining business during booms and depressions.
3. A careful study of cyclical variations facilitates businessman to face the recession period and make them ready to reap the benefits during booms.

4. Irregular Variation

Irregular Variation is the fourth type of change in time series analysis. In many situations, the value of variable may be completely unpredictable, changing in a random manner. Irregular variations describe such movements. The effect of the middle ease conflict in 1973 the ironian situation in 1979-1981 the collapse of OPEC in 1986 and the Iraqi situation in 1990 on gasoline prices in the united states are example of irregular variation.



3. 12 METHODS OF ESTIMATING TREND

There are four methods of estimating trend

1. The Free hand or Graphic Methods,
2. Semi - Average method,
3. Moving average method and
4. Method of Least Squares.

1. The Free hand or Graphic Methods

In this method we must plot the original data on the graph. Draw a smooth curve carefully, which will show the direction of the trend. To get proper trend line, we must note some points, while fitting a trend line by the free-hand method.

1. The curve should be smooth.
2. Approximately there must be equal number of points above and below the curve.
3. The total deviations of the data above the trend line must be the same as the vertical deviations below the line.

Space for Hint

4. The sum of the squares of the vertical deviations from the trend should be as small as possible.

Merits

1. It is the simplest, easiest, and quickest method. It saves time and labour.
2. It is adaptable and flexible, and it can be used to describe all types of trend (i.e) linear and non-linear.
3. Experienced statisticians can draw a free-hand line more accurately than a mathematician; this can widely be used in applied situations.
4. It will help to understand the character of time series, and we can use the appropriate mathematical trend.

Demerits

1. It is highly subjective. It is subject to personal bias. The result depends upon the judgment of the person who draws the line. There may be different curves for different persons.
2. It seems to be very simple. But it requires more time for a careful job.
3. If experienced persons do not draw it, then it is dangerous to use for forecasting purposes.
4. It does not help us to measure trend.

2. Semi - Average method.

In this method the original data are divided into two equal parts and averages are calculated for both the parts. These averages are called Semi – Averages.

Merits

1. It is simple and easier to understand than the moving average and least square method.
2. As it does not depend upon personal judgment, everyone who applies this method will get the same trend line unlike the former method.
3. As the line can be extended both ways, we can get the intermediate values and predict the future values.

Demerits

1. Under this method, it has an assumption of linear trend whether such a relationship exists or not.

2. It is affected by the limitation of arithmetic mean.

3. This method is not enough for forecasting the future trend or for removing trend from original data.

Thus moving average is better than the semi-average to estimate secular trend.

Example 3. 25

Find a trend line by the method of semi averages

Year	2003	2004	2005	2006	2007	2008
Sales in Units ('000)	60	77	82	120	116	130

Solution :

Let Y denotes the sales in units

Year	Sales in Units ('000)	semi moving total	semi moving average
2003	60		
2004	77	219	73
2005	82		
2006	120		
2007	116	366	122
2008	130		

3. Moving average method

In this method, the average value for a number of years or months or weeks is taken in to account and placing it at the centre of the time – span (period of moving average) and it is the normal or trend value for the middle period.

Space for Hint

Space for Hint

Calculation of moving averages

The formula for calculating 3 yearly moving average is

$$\frac{a+b+c}{3}, \frac{b+c+d}{3}, \frac{c+d+e}{3}$$

and the formula for 4 yearly moving average is

$$\frac{a+b+c+d}{4}, \frac{b+c+d+e}{4}, \frac{c+d+e+f}{4}$$

Merits

1. It is simple and easy to understand. It is easier to adopt when compared to the method of least square.
2. It is more flexible than other methods. It is elastic in the sense, items can be increased or decreased without affecting the moving average, the only Snag is that we get more trend values or less trend value.
3. It is not only used for the measurement of trend, but also for the measurement of seasonal, cyclical and irregular fluctuations.
4. The period of moving average is determined by the data and not by the personal judgment of the investigator. So is away from personal bias.

Demerits

1. In this method we cannot get the trend values for all the given observations. In finding trend value we leave the first and the last year in three-yearly moving average; and we leave the first two and the last two years in five yearly moving average.
2. The main object of trend value is that it is used for forecasting or predicting future values, because this method is not represented by a mathematical function.
3. There is no rule regarding the choice of the number of the moving average, and so the statistician has to use his own judgment.
4. In most of the economic and business time series the trend is a non-linear one; then the moving average lies below or above the curve of the actual data.

Due to these limitations, this method can be used when

1. The trend is linear is linear or approximately linear.
2. The oscillatory movements are regular, both in period and amplitude.
3. The future forecasting and current analysis is not required.

Example 3. 26

Find the 3 yearly moving average from the following time series data

Year	1998	1999	2000	2001	2002	2003	2004	2005
Sales in tons	30.1	45.4	39.3	41.4	42.2	46.4	46.6	49.2

Solution :

year	sales in tons	3 Year moving total	3 year moving average
1998	30.1		
1999	45.4	114.8	38.27
2000	39.3	126.1	42.03
2001	41.4	122.9	40.97
2002	42.2	130	43.33
2003	46.4	135.2	45.07
2004	46.6	142.2	47.40
2005	49.2		

Example 3. 27 :

Assuming a four-yearly cycle calculate the trend by the method of moving averages from the following data relating to the production of tea in India.

Year	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Sales	464	515	518	467	502	540	557	571	586	612

Space for Hint

Space for Hint

Solution :

Year	Sales	4 yearly total	Combined total	4 yearly Moving average
1991	464			
		1964		
1992	515		3966	495.75
		2002		
1993	518		4029	503.63
		2027		
1994	467		4093	511.63
		2066		
1995	502		4236	529.50
		2170		
1996	540		4424	553.00
		2254		
1997	557		4580	572.50
		2326		
1998	571			
1999	586			
2000	612			

4. Method of Least Squares

It is a mathematical as well as an analytical method. Under this method a straight-line trend can be fitted to the given time series of data. This is a most important and accurate method to measures the long-term trend. In this method we can fit either a straight line or a curve, which is the lies of best fit for the given data. By making use of the equation of this line we get the trend values. For all the given years we can predict or forecast the future trend values. The equation of the line of best fit can be taken as $y = ax + b$

where y denotes for the variable, x denotes the time period (which) is deviation from a particular year) and a, b are constants to be found out using the following two normal equations.

$$\Sigma y = a\Sigma x + nb \quad \text{and} \quad \Sigma xy = a\Sigma x^2 + b\Sigma x$$

Note : If we take x as the deviation from the middle year then $\Sigma x = 0$.

Then the normal equations becomes

$\Sigma y = nb$ and $\Sigma xy = a\Sigma x^2$

Thus $b = \frac{\Sigma y}{n}$ and $a = \frac{\Sigma xy}{\Sigma x^2}$

Substituting the values of a and b we get the equation of the straight-line trend.

Example 3. 28

Fit a straight line trend for the following data

Year	1992	1993	1994	1995	1996	1997	1998
Production in tons	62	70	75	81	89	93	104

Also find the trend values for the given years and estimate the trend value for the year 2001.

Solution :

Let the equation of the line of best fit be $y = ax + b$ ------------------(3.1)

To find a and b we use the following normal equations

$a\Sigma x_i + nb = \Sigma y_i$ ------------------------------------(3.2)

and $a\Sigma x_i^2 + b\Sigma x_i = \Sigma x_i y_i$ ---------------------------------- (3.3)

Let $X = \frac{x-1995}{1}$ and $Y = y - 89$

$\therefore$ (3.1) $\Rightarrow$ $Y = aX + b$ ---------------------------------- (3.4)

Space for Hint

Thus the normal equations (3.2) and (3.3) changed to

$$a\sum X_i + nb = \sum Y_i \qquad \text{-----------------------------------} (3.5)$$

and $a\sum X_i^2 + b\sum X_i = \sum X_i Y_i$ ----------------------------------- (3.6)

Now

x	y	$X = x-1995$	$Y = y-89$	x^2	xy	trend values
1992	62	-3	-27	9	81	62.08
1993	70	-2	-19	4	38	68.72
1994	75	-1	-14	1	14	75.36
1995	81	0	-8	0	0	82
1996	89	1	0	1	0	88.64
1997	93	2	4	4	8	95.28
1998	104	3	15	9	45	101.92
total		0	-49	28	186	

Thus (3.5) $\Rightarrow$ $7b = -49$

(i.e) $b = -7$

and (3.6) $\Rightarrow$ $28a = 186$

(i.e) $a = 6.64$

Now (3.4) $\Rightarrow$ $Y = 6.64X - 7$

Hence the required equation of line is $y = 6.64x - 13164.8$

The trend values are shown the last column of the table.

Example 3. 29

Using the method of least squares obtain the trend values for the following data

Year	2000	2001	2002	2003	2004	2005	2006	2007
Cost (Rs. in Lakhs)	43	41	38	39	32	35	27	25

Solution :

Let the equation of the line of best fit be $y = ax + b$ ------------------(3.7)

To find a and b we use the following normal equations

$a\sum x_i + nb = \sum y_i$ ------------------------------------(3.8)

and $a\sum x_i^2 + b\sum x_i = \sum x_i y_i$ ---------------------------------- (3.9)

Let $X = \dfrac{x - 2003.5}{0.5}$ and $Y = y - 39$

$\therefore$ (3.7) $\Rightarrow$ $Y = aX + b$ ---------------------------------- (3.10)

Thus the normal equations (3.8) and (3.9) changed to

$a\sum X_i + nb = \sum Y_i$ -----------------------------------(3.11)

and $a\sum X_i^2 + b\sum X_i = \sum X_i Y_i$ ---------------------------------- (3.12)

Now

x	y	$X = \dfrac{x-2003.5}{0.5}$	$Y = y - 39$	x^2	xy	trend values
2000	43	-7	4	49	-28	110
2001	41	-5	2	25	-10	102
2002	38	-3	-1	9	3	94
2003	39	-1	0	1	0	86
2004	32	1	-7	1	-7	78
2005	35	3	-4	9	-12	70
2006	27	5	-12	25	-60	62
2007	25	7	-14	49	-98	54
Total		0	-32	168	-212	

Space for Hint

Space for Hint

Thus (3.11) $\Rightarrow$ $8b = -32$

(i.e) $b = -4$

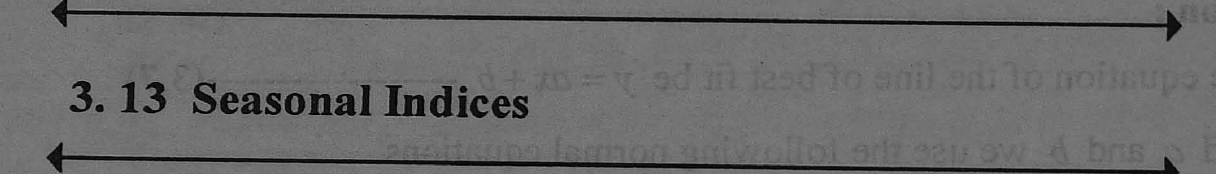
and (3.12) $\Rightarrow$ $168a = -212$

(i.e) $a = -1.26$

Now (3.10) $\Rightarrow$ $Y = -4X - 1.26$

Hence the required equation of line is $y = 8051.74 - 4x$

The trend values are shown the last column of the table.



3. 13 Seasonal Indices

In a time series, seasonal variations come into force in regular periods. There is monthly or quarterly seasonal variations in the economic and business phenomena. One of the methods is to find seasonal indices is methods of simple average.

Method of Simple Average

This is the simplest and easiest method of calculating a seasonal index the steps are

1. Average the data for each month or quarter for all the years.
2. Find the totals of each month or quarter.
3. Divide each total by the number of years for which data are given. If we are given monthly data for 4 years, we must first get the total for each month for 4 years and divide each total by 4 to get an average.
4. We can get an average of monthly averages by dividing the total of monthly averages by 12.
5. We must taken the average of month averages as 100 and get the seasonal Index as follows

$$\textit{Seasonal Index for January} = \frac{\textit{Monthly average for January} \times 100}{\text{Average of monthly average}}$$

Example 3. 30

Compute the average seasonal movement seasonal index for the following series.

Year	I quarter	II quarter	III quarter	IV quarter
1990	40	35	38	40
1991	42	37	39	38
1992	41	35	38	40
1993	45	36	36	41
1994	44	38	38	42

Solution :

Year	I quarter	II quarter	III quarter	IV quarter	Total
1990	40	35	38	40	
1991	42	37	39	38	
1992	41	35	38	40	
1993	45	36	36	41	
1994	44	38	38	42	
total	212	181	189	201	
Average	42.4	36.2	37.8	40.2	156.6
Seasonal index	108.30	92.46	96.55	102.68	39.15

$$Quarterly\ Average = \frac{quarterly\ total}{No.of\ seasons}$$

$$Seasonal\ Index = \frac{quarterly\ average}{General\ Average} \times 100$$

$$= \frac{156.6}{4}$$

$$= 39.15$$

Space for Hint

Space for Hint

Merits and Demerits

It is the simplest and the easiest method; but the seasonal index have the trend influence also. This is useful only when the original data have short-term tendencies.

Check Your Progress

(1) Compute the average seasonal movement seasonal index for the following series.

Year	Summer	Monsoon	Autumn	Winter
1996	68	60	61	63
1997	70	58	56	60
1998	68	63	68	67
1999	65	56	56	55
2000	60	55	55	58

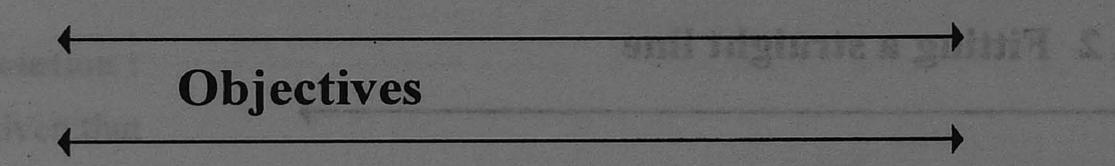
SUMMARY

In this unit we have learned the meaning of index number and their uses. Also we have discussed how to analyze the time series. Further we have discussed that how to find index numbers and seasonal indices.

Unit IV

Space for Hint

Curve Fitting



Objectives

In this unit, we are going to discuss how to find the equation of a curve which pass through maximum number of points.

After the completion of this unit one may able to fit

- a straight line
- a second degree parabola
- a curve $y = ab^x$

Introduction

Let x be an independent variable having values $x_1, x_2, x_3, \cdots, x_n$ and the corresponding dependent variable y having the values $y_1, y_2, y_3, \cdots, y_n$.

If the points (x_i, y_i), $i = 1,2,3,\cdots,n$ are plotted on a graph, the corresponding diagram is called a *scatter* diagram. Curve fitting is a process to find the functional relationship between the variables x and y.

4. 1 Principles of least squares

For the curve fitting the most familiar and easy method is principle of least squares method.

Let x be the independent variable having the values $x_1, x_2, x_3, \cdots, x_n$ and $y_1, y_2, y_3, \cdots, y_n$ be the respective values of the dependent variable y.

Let $y = f(x)$ be the equation of the required curve.

Let $d_i = y_i - f(x_i)$ where y_i is the given value of y called observed value and $f(x_i)$ is the value of y obtained from the functional relation.

Space for Hint

The value of d_i is called residual. The principle of least squares is to select $y = f(x)$ in such a way that $\sum^{n} d_i^2$ is minimum.

4. 2 Fitting a straight line

Let x have the values $x_1, x_2, x_3, \cdots, x_n$ and y have the values $y_1, y_2, y_3, \cdots, y_n$.

Let $y = ax + b$ be the equation of a straight line which pass through maximum number of points of the scatter diagram.

Now the residual d_i is given by $d_i = y_i - (ax_i + b)$.

Let $R = \Sigma d_i^2$

(i.e) $R = \Sigma[y_i - ax_i - b]^2$

Now according to the principle of least squares, we have to find a and b such that R is minimum.

For that, we shall find the partial derivatives of R with respect to a and b, and then equate to zero.

Now $\dfrac{\partial R}{\partial a} = 0$

$\Rightarrow -2\Sigma[y_i - ax_i - b]x_i = 0$

$\Rightarrow \Sigma[y_i - ax_i - b]x_i = 0$

$\Rightarrow \Sigma x_i y_i = a\Sigma x_i^2 + b\Sigma x_i$ --- (4.1)

and $\dfrac{\partial R}{\partial b} = 0$

$\Rightarrow -2\Sigma[y_i - ax_i - b] = 0$

$\Rightarrow \Sigma[y_i - ax_i - b] = 0$

$\Rightarrow \Sigma y_i = a\Sigma x_i + nb$ --- (4.2)

Solving (4.1) and(4.2), we get the values of a and b.

Note : Equations (4.1) and (4.2) are called normal equations of the line of best fit of $y = ax + b$

$\Sigma y_i = a\Sigma x_i + nb$.

Space for Hint

Example 4. 1 :

Fit a straight line to the following data:

x	:	0	1	2	3	4
y	:	1	1.8	3.3	4.5	6.3

Solution :

Given that

x	:	0	1	2	3	4
y	:	1	1.8	3.3	4.5	6.3

Let the line of best fit be $y = ax + b$ ---------------------------------- (4.3)

To find a and b we use the following normal equations

$a\sum x_i + nb = \sum y_i$ ---(4.4)

and $a\sum x_i^2 + b\sum x_i = \sum x_i y_i$ --------------------------------------- (4.5)

Now

	x	y	x^2	xy
	0	1	0	0
	1	1.8	1	1.8
	2	3.3	4	6.6
	3	4.5	9	13.5
	4	6.3	16	25.2
total	10	16.9	30	47.1

Thus (4.4) $\Rightarrow$ $10a + 5b = 16.9$ --------------------------------------(4.6)

and (4.5) $\Rightarrow$ $30a + 10b = 47.1$ --(4.7)

Solving (4.6) and (4.7), we get, $a = 1.33$ and $b = 0.72$

Hence the required equation of line is $y = 1.33x + 0.72$

Example 4. 2

Fit a straight line to the following data:

Years	:	1911	1921	1931	1941	1951
Production in (tons)	:	10	12	8	10	14

Space for Hint

Solution :

Given that

Years	:	1911	1921	1931	1941	1951
Production in (tons)	:	10	12	8	10	14

Let the line of best fit be $y = ax + b$ ---------------------------------- (4.8)

Let $X = \dfrac{x - 1931}{10}$ and $Y = y$

Thus (4.8) becomes $Y = aX + b$ ---------------------------------- (4.9)

To find a and b we use the following normal equations

$a\sum X_i + nb = \sum Y_i$ ---------------------------------- (4.10)

and $a\sum X_i^2 + b\sum X_i = \sum X_i Y_i$ ---------------------------------- (4.11)

Now

	x	y	X	Y	X^2	XY
	1911	10	-2	10	4	-20
	1921	12	-1	12	1	-12
	1931	8	0	8	0	0
	1941	10	1	10	1	10
	1951	14	2	14	4	28
Total			**0**	**54**	**10**	**6**

Thus (4.10) $\Rightarrow a(0) + 5b = 54$

(i.e) $b = 10.8$

and (4.11) $\Rightarrow 10a + b(0) = 6$

(i.e) $a = 0.6$

Hence the required equation of line is $Y = 0.6X + 10.8$

(i.e) $y = 0.6\left(\dfrac{x - 1931}{10}\right) + 10.8$

(i.e) $y = 0.06x + 126.66$

Space for Hint

Check Your Progress

(1) Fit a straight line to the following data :

x	0	1	2	3	4
y	2.1	3.5	5.4	7.3	8.2

(Answer : $y = 1.6x + 2.1$)

(2) Fit a straight line to the following data :

x	0	5	10	15	20	25
y	12	15	17	22	24	30

(Answer : $y = 0.698x + 11.27$)

4. 3 Fitting a Second degree parabola

Let x have the values $x_1, x_2, x_3, \cdots, x_n$ and y have the values $y_1, y_2, y_3, \cdots, y_n$.

Let $y = ax^2 + bx + c$ be the required equation of a second degree parabola.

Now the residual d_i is given by $d_i = y_i - (ax_i^2 + bx_i + c)$.

Let $R = \Sigma d_i^2$

(i.e) $R = \Sigma[y_i - ax_i^2 - bx_i - c]^2$

Now by the principle of least squares, we have to find a, b and c such that R is minimum.

For that, we shall find the partial derivatives of R with respect to a, b and c, and then equate to zero.

Now $\dfrac{\partial R}{\partial a} = 0$

$\Rightarrow -2\Sigma[y_i - ax_i^2 - bx_i - c]x_i^2 = 0$

$\Rightarrow \Sigma x_i^2 y_i - \Sigma a x_i^4 - b\Sigma x_i^3 - c\Sigma x_i^2 = 0$

$\Rightarrow \Sigma a x_i^4 + b\Sigma x_i^3 + c\Sigma x_i^2 = \Sigma x_i^2 y_i$ ---------------------------------- (4.12)

Space for Hint

and $\frac{\partial R}{\partial b} = 0$

$\Rightarrow -2\sum[y_i - ax_i^2 - bx_i - c]x_i = 0$

$\Rightarrow \sum x_i y_i - \sum ax_i^3 - b\sum x_i^2 - c\sum x_i = 0$

$\Rightarrow \sum ax_i^3 + b\sum x_i^2 + c\sum x_i = \sum x_i y_i$ ---------------------------------- (4.13)

and $\frac{\partial R}{\partial c} = 0$

$\Rightarrow -2\sum[y_i - ax_i^2 - bx_i - c] = 0$

$\Rightarrow \sum y_i - \sum ax_i^2 - b\sum x_i - nc = 0$

$\Rightarrow \sum ax_i^2 + b\sum x_i + nc = \sum y_i$ -- (4.14)

Solving (4.12), (4.13) and (4.14), we get the values of a, b and c.

Note : Equations (4.12), (4.13) and (4.14) are called normal equations to fit second degree parabola $y = ax^2 + bx + c$

Example 4.3 :

Fit a second degree parabola to the following data:

x	:	0	1	2	3	4
y	:	1	1.8	1.3	2.5	2.3

Solution :

Given that

x	:	0	1	2	3	4
y	:	1	1.8	1.3	2.5	2.3

Let the equation of second degree parabola be $y = ax^2 + bx + c$ -- (4.15)

To find a, b and c we use the following normal equations

$\sum ax_i^2 + b\sum x_i + nc = \sum y_i$ ------------------------------------ (4.16),

$\sum ax_i^3 + b\sum x_i^2 + c\sum x_i = \sum x_i y_i$ ------------------------------ (4.17)

and $\sum ax_i^4 + b\sum x_i^3 + c\sum x_i^2 = \sum x_i^2 y_i$ ---------------------------- (4.18)

Now

	x	y	x^2	x^3	x^4	xy	x^2y
	0	1	0	0	0	0	0
	1	1.8	1	1	1	1.8	1.8
	2	1.3	4	8	16	2.6	5.2
	3	2.5	9	27	81	7.5	22.5
	4	2.3	16	64	256	9.2	36.8
total	10	8.9	30	100	354	21.1	66.3

Thus (4.16) $\Rightarrow$ $30a + 10b + 5c = 8.9$ ---------------------------------(4.19)

and (4.17) $\Rightarrow$ $100a + 30b + 10c = 21.1$ ----------------------------(4.20)

and (4.18) $\Rightarrow$ $354a + 100b + 30c = 66.3$ -------------------------(4.21)

Solving (4.19), (4.20) and (4.21), we get, $a = -0.021$, $b = 0.416$ and $c = 1.077$.

Hence the required equation of parabola is $y = -0.021x^2 + 0.416x + 1.077$.

Example 4. 4 :

Fit a second degree parabola to the following data:

x	2	4	6	8	10	12
y	5.5	9.1	14.9	22.8	33.3	46

Solution :

Given that

x	2	4	6	8	10	12
y	5.5	9.1	14.9	22.8	33.3	46

Let the equation of second degree parabola be $y = ax^2 + bx + c$ -- (4.22)

To find a, b and c we use the following normal equations

$\sum ax_i^2 + b\sum x_i + nc = \sum y_i$ ------------------------------------(4.23),

$\sum ax_i^3 + b\sum x_i^2 + c\sum x_i = \sum x_i y_i$ ------------------------------ (4.24)

and $\sum ax_i^4 + b\sum x_i^3 + c\sum x_i^2 = \sum x_i^2 y_i$ ----------------------------- (4.25)

Space for Hint

Space for Hint

Now

	x	y	x^2	x^3	x^4	xy	x^2y
	2	5.5	4	8	16	11	22
	4	9.1	16	64	256	36.4	145.6
	6	14.9	36	216	1296	89.4	536.4
	8	22.8	64	512	4096	182.4	1459.2
	10	33.3	100	1000	10000	333	3330
	12	46	144	1728	20736	552	6624
total	42	131.6	364	3528	36400	1204.2	12117.2

Thus (4.23) $\Rightarrow 364a + 42b + 6c = 131.6$ -------------------------- (4.26)

and (4.24) $\Rightarrow 3528a + 364b + 42c = 1204.2$ -------------------- (4.27)

and (4.25) $\Rightarrow 36400a + 3528b + 364c = 12117.2$ ------------ (4.28)

Solving(4.26), (4.27) and (4.28), we get, $a = 0.287$, $b = 0.024$ and $c = 4.35$.

Hence the required equation of parabola is $y = 0.287x^2 + 0.024x + 4.35$.

Check Your Progress

(1) Fit a second degree parabola to the following data:

x	0	10	20	30	40	50
y	115	160	215	270	335	400

(Answer : $y = 0.025x^2 + 4.4786x + +114.285$)

(2) Fit a second degree parabola to the following data:

x	0	2	4	6	8	10
y	1	3	13	32	57	91

(Answer : $y = 0.9821x^2 - 0.8071x + 0.8571$)

Space for Hint

4. 4 Fitting a curve of the form $y = be^{ax}$

Let x have the values $x_1, x_2, x_3, \cdots, x_n$ and y have the values $y_1, y_2, y_3, \cdots, y_n$.

Now we want to fit the curve $y = bx^a$ ------------------------ (4.29)

Taking logarithms on both sides of (4.29), we have, $\log y = \log b + a \log x$

(i.e) $Y = AX + B$ -- (4.30)

where $X = \log x$, $Y = \log y$ and $B = \log b$.

Clearly (4.30) is linear in X and Y, and hence to find A and B, we use the following normal equations

$A\sum X_i + nB = \sum Y_i$ -- (4.31)

and $\sum X_i Y_i = A\sum X_i^2 + B\sum X_i$ -- (4.32)

Solving (4.31), and (4.32), we get the values of A and B.

Thus we able to find the values of a and b

Example 4. 5 :

Fit a curve $y = bx^a$ to the following data:

x	:	0	1	2	3	4
y	:	1	1.8	1.3	2.5	2.3

Solution :

Given that

x	:	0	1	2	3	4
y	:	1	1.8	1.3	2.5	2.3

To fit $y = bx^a$ the normal equations are

$A\sum X_i + nB = \sum Y_i$ -- (4.33)

and $\sum X_i Y_i = A\sum X_i^2 + B\sum X_i$ -- (4.34)

Space for Hint

x	y	X	Y	X^2	XY
1	2.99	0.0000	0.4757	0.0000	0.0000
2	4.25	0.3010	0.6284	0.0906	0.1892
3	5.22	0.4771	0.7177	0.2276	0.3424
4	6.1	0.6021	0.7853	0.3625	0.4728
Total		1.3802	2.6071	0.6807	1.0044

Thus (4.33) $\Rightarrow 1.3802A + 4B = 2.0671$ -------------------------- (4.35)

and (4.34) $\Rightarrow 0.6807A + 1.3802B = 1.0044$ --------------------- (4.36)

Solving (4.35) and (4.36), we get, $A = 0.5126$ and $B = 0.4749$

Thus $a = 0.5126$ and $b = 2.9847$

Hence the required curve is $y = 2.9847x^{0.5126}$

Example 4. 6 :

Fit a curve $y = a \cdot e^{bx}$ to the following data:

x	:	2	3	4	5	6
y	:	144	172.8	207.4	248.8	298.5

Solution :

Given that

x	:	2	3	4	5	6
y	:	144	172.8	207.4	248.8	298.5

Taking natural logarithms on both sides, we have, $y = a \cdot e^{bx}$

$\ln y = \ln a + bx$

(i.e) $Y = A + BX$ --- (4.37)

where $X = x$, $Y = \ln y$, $A = \ln A$ and $B = b$.

To find the values of A and B, the normal equations are

$B\sum X_i + nA = \sum Y_i$ --- (4.38)

and $B\sum X_i^2 + A\sum X_i = \sum X_i Y_i$ ------------------------------------- (4.39)

x	y	X	Y	X^2	XY
2	144	0.6931	4.9698	0.4805	3.4448
3	172.8	1.0986	5.1521	1.2069	5.6602
4	207.4	1.3863	5.3346	1.9218	7.3954
5	248.8	1.6094	5.5166	2.5903	8.8787
6	298.5	1.7918	5.6988	3.2104	10.2108
total		6.5793	26.6720	9.4099	35.5899

Thus (4.38) $\Rightarrow 6.5793B + 5A = 26.6720$ ------------------------ (4.40)

and (4.39) $\Rightarrow 9.4099B + 6.5793A = 35.5899$ ------------------- (4.41)

Solving (4.40) and (4.41), we get, $A = 4.4718$ and $B = 0.6556$

Thus $a = 87.5122$ and $b = 0.6556$

Hence the required curve is $y = 87.5122x^{0.6556}$

Check Your Progress

(1) Fit the exponential curve $y = ae^{bx}$ to the following data.

No. of petals	5	6	7	8	9	10
No. of flowers	133	55	23	7	2	2

(2) Fit the exponential curve $y = ae^{bx}$ to the following data.

x	1	2	3	4
y	2.99	4.25	5.22	6.10

Space for Hint

Space for Hint

(3) Fit a second degree parabola to the following data.

x	1	2	3	4	5	6	7
y	2.3	5.2	9.7	16.5	29.4	35.5	54.4

(4) Fit a second degree parabola to the following data.

x	10	15	20	25	30	35	40
y	11	13	16	20	27	34	41

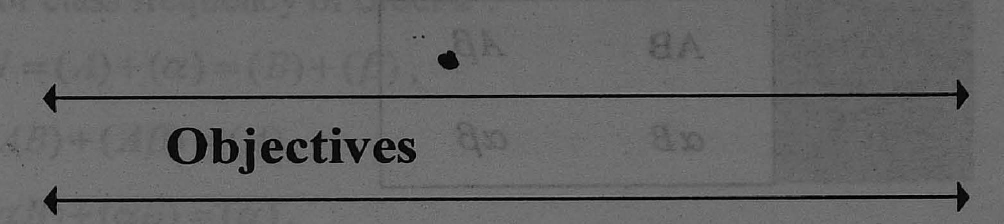
SUMMARY

In this unit we have learned that how to fit a straight line, second degree parabola and exponential curves.

Unit V

Space for Hint

Attributes



Objectives

In this unit, we are going to discuss how to find the relationship between two attributes, independency of two attributes.

After the completion of this unit one may able to find the

- Relationship between attributes
- Independency of two attributes

Introduction

In the earlier units we have discussed those data related to quantitative. But we cannot measure those data having qualitative nature. Qualitative characteristics of a population are called attributes and they cannot be measured by numeric quantity.

Suppose a population is divided into two classes according to the possession or non-possession or presence or absence of a single attribute. The class in which a particular characteristics is present is called positive class and it is denoted by upper case of English alphabets and the absence of the characteristics is called negative class and it denoted by the lower case of Greek letters. That is $A, B, C, \cdots$ denotes the presence of characteristics and $\alpha, \beta, \gamma, \cdots$ denotes the corresponding absence of the characteristics.

For example, if attribute A represents *rich* and B represents *literate* then α refers *poor* and β refers *illiterate*. Further AB represents the possession of both rich and literate; $A\beta$ represents rich and illiterate; αB represents poor and literate and $\alpha\beta$ represents poor and illiterate.

The above example can be represent as table form as

Space for Hint

Attribute	B	β
A	AB	$A\beta$
α	αB	$\alpha\beta$

A class represented by n attributes is called a class of n^{th} order. Thus A, B, α, β are called first order, $AB, A\beta, \alpha B, \alpha\beta$ are called class of second order, $ABC, AB\gamma, A\beta\gamma, \cdots$ are called class of third order.

Now the number of individuals possess the attribute A is called frequency of the attribute A and it is denoted by (A). Hence (αB) stands the number of individuals possessing the attributes α and B.

The total frequency in a population is denoted by N.

Note :

Class frequencies of the type $(A), (AB), (ABC), \cdots$ are called positive class frequencies.

Class frequencies of the type $(\alpha), (\alpha\beta), (\alpha\beta\gamma), \cdots$ are called negative class frequencies.

Class frequencies of the type $(A\beta), (\alpha B), (A\beta C), \cdots$ are called contrary class frequencies.

5. 1 Attributes

The frequency classes for two attributes can be represented in the form of a table which is given below.

Attribute	B	β	Total
A	AB	$A\beta$	(A)
α	αB	$\alpha\beta$	(α)
Total	(B)	(β)	N

Space for Hint

The relationship between the class frequencies of various orders are given below.

For class frequency of order 2 :

$N=(A)+(\alpha)=(B)+(\beta)$,

$(AB)+(A\beta)=(A)$,

$(\alpha B)+(\alpha\beta)=(\alpha)$,

$(AB)+(\alpha B)=(B)$,

$(A\beta)+(\alpha\beta)=(\beta)$,

$(AB)+(A\beta)+(\alpha B)+(\alpha\beta)=N$.

For class frequency of order 3 :

$(ABC)+(AB\gamma)+(A\beta C)+(A\beta\gamma)=(A)$,

$(ABC)+(A\beta\gamma)+(\alpha BC)+(\alpha B\gamma)=(B)$,

$(ABC)+(A\beta C)+(\alpha BC)+(\alpha\beta C)=(C)$,

$(ABC)+(AB\gamma)=(AB)$,

$(ABC)+(A\beta\gamma)=(A\gamma)$ etc.,

Theorem 5. 1

For n attributes

(a) total number of class frequencies is 3^n

(b) total number of positive class frequencies is 2^n

(c) total number of negative class frequencies is 2^n-1.

Proof :

Let n attributes be given.

Thus the number of ways of choosing r attributes from the given set of n attributes is $\binom{n}{r}$.

Since each attribute gives two symbols (one for positive class and another for negative class), the number of class frequencies of order r that can be obtained from r attributes is 2^r.

Hence the total number of class frequencies of order r is $\binom{n}{r}2^r$.

Space for Hint

Thus the total number of all class frequencies = $\sum_{r=0}^{n}\binom{n}{r}2^r$

$$= 1+\binom{n}{1}2^1+\binom{n}{2}2^2+\binom{n}{3}2^3+\cdots+\binom{n}{n}2^n$$

$$= (1+2)^n$$

$$= 3^n$$

(ii) Further any collection of r attributes have only one positive class frequency of order r.

Hence the total number of positive class frequencies of all orders

$$= \sum_{r=0}^{n}\binom{n}{r}$$

$$= 1+\binom{n}{1}+\binom{n}{2}+\binom{n}{3}+\cdots+\binom{n}{n}$$

$$= (1+1)^n$$

$$= 2^n$$

(iii) clearly there is no negative class frequency of order 0.

Thus any collection of r attributes gives rise to one negative class frequency of order r.

Hence the total number of negative class frequencies of all orders

$$= \sum_{r=0}^{n}\binom{n}{r}-1$$

$$= 2^n-1.$$

Dichotomization :

Dichotomization is the process of dividing a collection of objects into two classes according to the possession or non-possession of an attribute.

Notations :

We use the following notations in the theory of attribute.

The total number of objects in a population is N

If A is an attribute then $N=(A)+(\alpha)$ and it is written as $(A)=A\cdot N$ and $(\alpha)=\alpha\cdot N$.

Thus $N=(A)+(\alpha)$

$$= A\cdot N+\alpha\cdot N$$

$$= (A+\alpha)\cdot N$$

Hence $\boxed{A+\alpha=1}$.

Thus in symbolic expression A can be replaced by $1-\alpha$ and α by $1-A$.

Example 5. 1

Prove that $(BC)=(ABC)+(\alpha BC)$

Proof :

Now $(\alpha BC) = \alpha BC\cdot N$

$$= (1-A)BC\cdot N$$

$$= BC\cdot N-ABC\cdot N$$

$$= (BC)-(ABC)$$

Hence $(BC)=(ABC)+(\alpha BC)$.

Example 5. 2

Prove that for two attributes A and B, prove that

$N=(AB)+(A\beta)+(\alpha B)+(\alpha\beta)$.

Proof :

Let A and B be two attributes.

$\therefore\ N = (A)+(\alpha)$

$$= (AB)+(A\beta)+(\alpha B)+(\alpha\beta)$$

Hence $N=(AB)+(A\beta)+(\alpha B)+(\alpha\beta)$

Example 5. 3

For any three attributes, prove that

$N=(ABC)+(AB\gamma)+(A\beta C)+(A\beta\gamma)+(\alpha BC)+(\alpha B\gamma)+(\alpha\beta C)+(\alpha\beta\gamma)$.

Proof :

L.H.S. = N

$$= (AB)+(A\beta)+(\alpha B)+(\alpha\beta)$$

$$= (ABC)+(AB\gamma)+(A\beta C)+(A\beta\gamma)+(\alpha BC)+(\alpha B\gamma)+(\alpha\beta C)+(\alpha\beta\gamma)$$

= R.H.S

Space for Hint

Space for Hint

Example 5. 4

Prove that for any two attributes negative class frequencies can be expressed in terms of positive class frequencies and converse also true.

Proof :

Consider two attributes A and B.

Now $(\alpha\beta) = \alpha\beta \cdot N$

$= (1-A)(1-B) \cdot N$

$= (1-A-B+AB) \cdot N$

$= N - A \cdot N - B \cdot N + AB \cdot N$

$= N-(A)-(B)+(AB)$

Hence $\boxed{(\alpha\beta) = N-(A)-(B)+(AB)}$

Thus negative class frequencies can be expressed in terms of positive class frequencies.

and $(AB) = AB \cdot N$

$= (1-\alpha)(1-\beta) \cdot N$

$= (1-\alpha-\beta+\alpha\beta) \cdot N$

$= N - \alpha \cdot N - \beta \cdot N + \alpha\beta \cdot N$

$= N-(\alpha)-(\beta)+(\alpha\beta)$

Hence $\boxed{(AB) = N-(\alpha)-(\beta)+(\alpha\beta)}$

Thus positive class frequencies can be expressed in terms of negative class frequencies.

Example 5. 5

Show that for n attributes $A_1, A_2, A_3, \cdots, A_n$,

$(A_1, A_2, A_3, \cdots, A_n) \geq (A_1)+(A_2)+(A_3)+\cdots+(A_n)-(n-1)N$.

Proof :

We shall prove the result using induction on n.

If $n=2$ then $(\alpha_1\alpha_2) = N-(A_1)-(A_2)+(A_1A_2)$

We know that $(\alpha_1\alpha_2) \geq 0$

(i.e.) $N-(A_1)-(A_2)+(A_1A_2) \geq 0$

(i.e.) $(A_1A_2) \geq (A_1)+(A_2)-N$

(i.e.) $(A_1A_2) \geq (A_1)+(A_2)-(2-1)N$

Thus the result is true for $n=2$.

Assume the result is true for $n=k$.

(i.e.) $(A_1, A_2, A_3, \cdots, A_k) \geq (A_1)+(A_2)+(A_3)+\cdots+(A_k)-(k-1)N$

Now we shall prove the result is true for $n=k+1$

Now $(A_1, A_2, \cdots, A_{k+1}) \geq (A_1)+(A_2)+\cdots+(A_{k-1})+(A_kA_{k+1})-(k-1)N$

(i.e.) $(A_1, A_2, \cdots, A_{k+1}) \geq (A_1)+(A_2)+\cdots+(A_{k-1})+(A_k)+(A_{k+1})-N-(k-1)N$

(i.e.) $(A_1, A_2, \cdots, A_{k+1}) \geq (A_1)+(A_2)+\cdots+(A_{k-1})+(A_k)+(A_{k+1})-kN$

(i.e.) $(A_1, A_2, \cdots, A_{k+1}) \geq (A_1)+(A_2)+\cdots+(A_{k-1})+(A_k)+(A_{k+1})-(\overline{k+1}-1)N$

(i.e.) the result is true for $n=k+1$

$\therefore$ by mathematical induction, the result is true for all positive integer n

(i.e.) for $n \in \mathbf{Z}^+$, $(A_1, A_2, \cdots, A_n) \geq (A_1)+(A_2)+\cdots+(A_n)-(n-1)N$.

This proves the problem.

Example 5. 6

Given frequencies $(A)=1150$, $(\alpha)=1120$, $(AB)=1075$, $(\alpha\beta)=985$. Find the remaining class frequencies and the total number of the observations.

Solution :

Given that $(A)=1150$, $(\alpha)=1120$, $(AB)=1075$, $(\alpha\beta)=985$.

We shall form the contingency table as follows.

Attribute	B	β	Total
A	1075	75	1150
α	135	985	1120
Total	1210	1060	2270

From the above table it is clear that $(A\beta)=75$, $(\alpha\beta)=135$, $(\beta)=985$, $(B)=1210$ and total number of observations = $N=2270$.

Space for Hint

Space for Hint

Example 5. 7

Given the following class frequencies, find the frequencies of the positive and negative classes and the total number of observations.

$(AB) = 733$, $(A\beta) = 840$, $(\alpha B) = 699$, $(\alpha\beta) = 783$.

Solution :

Given that $(AB) = 733$, $(A\beta) = 840$, $(\alpha B) = 699$, $(\alpha\beta) = 783$.

We shall form the contingency table as follows.

Attribute	B	β	Total
A	733	840	1573
α	699	783	1482
Total	1432	1623	3055

From the above table it is clear that,

negative class frequencies : $(\alpha) = 1482$ and $(\beta) = 1623$,

positive class frequencies : $(A) = 1573$, $(B) = 1432$

total number of observations = $N = 3055$.

Example 5. 8

A survey reveals that out of 1000 people in a locality 800 like coffee; 700 like tea; 660 like both coffee and tea. Find how many people like neither coffee nor tea.

Solution :

Let A be attribute to denote the coffee liker.

Let B be attribute to denote the tea liker.

We shall form the contingency table as follows.

Attribute	B	β	Total
A	660	140	800
α	40	160	200
Total	700	300	1000

From the above table, the number of people like neither coffee nor tea is $(\alpha\beta) = 160$.

Example 5. 9

100 children took three examinations. 40 passed the firs, 39 passed the second and 48 passed the third, 10 passed all three. 21 failed all three, 9 passed the first two and failed the third, 19 failed the first two and passed the third. Find how many children passed at least two examinations.

Solution :

Let A be attribute to denote a child passed in the first examination.

Let B be attribute to denote a child passed in the second examination.

Let C be attribute to denote a child passed in the third examination.

Given that $N = 100$, $(A) = 40$, $(B) = 39$, $(C) = 48$,

$(ABC) = 10$, $(\alpha\beta\gamma) = 21$, $(AB\gamma) = 9$, $(\alpha\beta C) = 19$.

We know that $(ABC) + (\alpha BC) + (A\beta C)$

$= (C) - (\alpha\beta C)$

$= 48 - 19$

$= 29$

$\therefore$ number of children passed in at least two examinations

$= (ABC) + (\alpha BC) + (A\beta C) + (AB\gamma)$

$= 29 + 9$

$= 38$

Space for Hint

Space for Hint

Example 5. 10

A survey conducted among T.V. viewers in a city revealed the following results. 850 see Doordharshan T.V. programmes; 780 set Star T.V. programmes; 326 see Cable T.V. programmes; 50 see all the three programmes; 200 see Doordharshan T.V. programme and Star T.V.programmes but not Cable T.V. programmes;110 do not see Doordharshan and Star T.V. programmes but see Cable T.V. programmes.

(i) Find how many people see Doordharshan and Star T.V. programmes

(ii) Find how many people see at least two T.V. programmes.

Solution :

Let A be attribute to denote those see Doordharshan T.V. programmes.

Let B be attribute to denote those see Star T.V. programmes.

Let C be attribute to denote those see Cable T.V. programmes.

Given that $(A) = 850$, $(B) = 70$, $(C) = 326$, $(ABC) = 200$, $(\alpha\beta C) = 21$.

(i) Now the number people see Doordharshan and Star T.V. programmes

$= (AB)$

$= (ABC) + (AB\gamma)$

$= 50 + 200$

$= 250.$

(ii) We know that $(ABC) + (\alpha BC) + (A\beta C)$

$= (C) - (\alpha\beta C)$

$= 326 - 110$

$= 216$

The number of people see at least two T.V. programmes

$= (ABC) + (\alpha BC) + (A\beta C) + (AB\gamma)$

$= 216 + 200$

$= 416$

Check Your Progress

(1) Of 500 men in a locality exposed to cholera 172 in all were attacked; 178 were inoculated and of these 128 were attacked. Find the number o persons

(a) not inoculated not attacked

(b) inoculated not attacked

(c) not inoculated attacked

(2) There were 200 students in a college whose results in the First semester, second semester and the third semester are as follows.

80 passed in the first semester; 75 passed in the second semester;

96 passed in the third semester; 25 passed in all the three semesters;

46 failed in all the three semesters; 29 passed in the first two and failed in the third semester;

42 failed in the first two semesters and passed in the third semester.

75 passed in the second semester;

Find how many students passed in at least two semesters.

(3) In a very hotly fought battle 70% of the solders at least lost an eye, 75% at least lost an ear; 80% at least an arm and 85% at least lost a leg. How many at least lost all the four.

(4) In a University examinations 95% of the candidates passed Part I, 70% passed Part II, 65% passed Part III. Find how many at least should have passed the whole examinations.

5. 2 Consistency of Data

Consider a population with the attribute A and B. For the data observed in the same population (AB) cannot be greater than (A). Hence the figures $(A)=200$ and $(AB)=250$ are inconsistent and $(\beta)=-800$ is also inconsistent

Definition : A set of class frequencies is said to be *consistent* if none of them is negative.

The following are a set of criteria for testing the consistency in the case of single attribute, two attributes and three attributes.

Space for Hint

Space for Hint

Attribute	Condition for consistency	Equivalent positive class condition	Number of conditions
A	$(A)\geq 0$ $(\alpha)\geq 0$	$(A)\geq 0$ $(A)\leq N$	2
A, B	$(AB)\geq 0$ $(A\beta)\geq 0$ $(\alpha B)\geq 0$ $(\alpha\beta)\geq 0$	$(AB)\geq 0$ $(AB)\leq (A)$ $(AB)\leq (B)$ $(AB)\geq (A)+(B)-N$	2^2
A, B, C	$(ABC)\geq 0$ $(AB\gamma)\geq 0$ $(A\beta C)\geq 0$ $(\alpha BC)\geq 0$ $(A\beta\gamma)\geq 0$ $(\alpha B\gamma)\geq 0$ $(\alpha\beta C)\geq 0$ $(\alpha\beta\gamma)\geq 0$	(1) $(ABC)\geq 0$ (2) $(ABC)\leq (AB)$ (3) $(ABC)\leq (AC)$ (4) $(\alpha BC)\leq (BC)$ (5) $(ABC)\geq (AB)+(AC)-(A)$ (6) $(ABC)\geq (AC)+(BC)-(C)$ (7) $(ABC)\geq (AC)+(BC)-(C)$ (8) $(ABC)\leq (AB)+(BC)+(AC)$ $-(A)-(B)-(C)+N$ (9) $(AB)+(BC)+(AC)\geq$ $(A)+(B)+(C)-N$ (10) $(AC)+(BC)-(AB)\leq (C)$ (11) $(AB)+(BC)-(AC)\leq (B)$ (12) $(AB)+(AC)-(BC)\leq (A)$	2^3

Example 5. 11

Test the consistency of the data when $(A)=800$, $(B)=700$, $(AB)=660$, $N=1000$

Solution :

For the given data, test the consistency.

$(A)=800$, $(B)=700$, $(AB)=50$, $N=1000$.

Solution :

Given that $(A)=800$, $(B)=700$, $(AB)=50$, $N=1000$.

We shall find the other class frequencies

Attribute	B	β	Total
A	660	140	800
α	40	160	200
Total	700	300	1000

Since all class frequencies are positive, therefore the give data is consistent.

Example 5. 12

Test the consistency of the data when

$(A) = 600$, $(B) = 500$, $(AB) = 50$, $N = 1000$.

Solution :

For the given data, test the consistency.

$(A) = 600$, $(B) = 500$, $(AB) = 50$, $N = 1000$.

Solution :

Given that $(A) = 600$, $(B) = 500$, $(AB) = 50$, $N = 1000$.

We shall find the other class frequencies

Attribute	B	β	Total
A	50	550	600
α	450	-50	400
Total	500	500	1000

Since the class frequency $(\alpha\beta)$ is negative, therefore the give data is inconsistent.

Space for Hint

Space for Hint

Example 5. 13

A market investigator returns the following data. Of 2000 people consulted 1754 liked chocolates, 1872 liked toffee and 572 liked biscuits, 676 liked chocolates and toffee, 286 liked chocolate and biscuits, 270 liked toffee and biscuits, 114 liked all the three. Verify that the information given by the investigator is consistent.

Solution :

Let A be the attribute that liking chocolate, B be the attribute that liking toffee and C be the attribute that liking biscuit.

Given that $N = 2000$, $(A) = 1754$, $(B) = 1872$, $(C) = 572$, $(AB) = 676$, $(AC) = 286$, $(BC) = 270$, $(ABC) = 114$.

We know that $(\alpha\beta\gamma) = N - (A) - (B) - (C) + (AB) + (AC) + (BC) - (ABC)$

$$= 2000 - 1754 - 1872 - 572 + 676 + 286 + 270 - 114$$

$$= -1080$$

Since the class frequency $(\alpha\beta\gamma)$ is negative, therefore the give data is inconsistent.

Example 5. 14

The following summary appears in a report on a survey covering 1000 fields. Scrutinize the numbers and point out if there is any mistake or misprint in them.

Manured fields	510
Irrigated fields	490
Fields growing improved varieties	427
Fields both irrigated and manured	189
Fields both manured and growing improved varieties	140
Fields both irrigated and growing improved varieties	85

Solution :

Let A be the attribute of Manured fields, B be the attribute of Irrigated fields and C be the attribute of Fields growing improved varieties.

Space for Hint

Given that $N=1000$, $(A)=510$, $(B)=490$, $(C)=427$, $(AB)=189$, $(AC)=85$, $(BC)=140$.

We know that $(\alpha\beta\gamma)=N-(A)-(B)-(C)+(AB)+(AC+(BC)-(ABC)$

$= 1000-510-490-427+189+140+85-(ABC)$

$= -13-(ABC)$

Now $(\alpha\beta\gamma)\geq 0$

(i.e.) $-13-(ABC)\geq 0$

(i.e.) $(ABC)\leq -13<0$

Since the class frequency (ABC) is negative, therefore there is a mistake in the data.

Example 5. 15

If $(A)=50$, $(B)=60$, $(C)=50$, $(A\beta)=5$, $(A\gamma)=20$ and $N=100$. Find the least and greatest values of (BC).

Solution :

Given that $(A)=50$, $(B)=60$, $(C)=50$, $(A\beta)=5$, $(A\gamma)=20$ and $N=100$.

Now $(AB)=(A)-(A\beta)$

$= 50-5$

$= 45$

and $(AC)=(A)-(A\gamma)$

$= 50-20$

$= 30$

Now $(AB)+(BC)+(AC)\geq(A)+(B)+(C)-N$

(i.e.) $45+(BC)+30\geq 50+60+50-100$

$\Rightarrow (BC)\geq -15$ --(5.1)

Again $(AB)+(AC)-(BC)\leq(A)$

$\Rightarrow 45+30-(BC)\leq 50$

$\Rightarrow (BC)\geq 25$ --(5.2)

Now $(AB)+(BC)-(AC)\leq(B)$

$\Rightarrow 45+(BC)-30\leq 60$

$\Rightarrow (BC)\leq 45$ --(5.3)

Space for Hint

Now $(AC)+(BC)-(AB)\le(C)$

$\Rightarrow\ 30+(BC)-45\le 50$

$\Rightarrow\ (BC)\le 65$ -- (5.4)

Form (5.1), (5.2), (5.3) and (5.4), we have $-13\le 25\le(BC)\le 45\le 65$

$\Rightarrow\ 25\le(BC)\le 45$.

Check Your Progress

(1) If $N=120,\ (A)=60,\ (B)=90,\ (C)=30,\ (BC)=15,\ (AC)=15$, find the limits between which (AB) lie.

(2) Find the least and greatest values of (ABC) if $(A)=50$, $(B)=60,(C)=80,(AB)=35$, $(AC)=45$ and $(BC)=42$.

(3) Show that there is some error in the following data : 50% of people are wealthy and healthy, 35% are wealthy but not healthy, 20% are healthy but wealthy.

(4) Of 2000 people consulted 1854 speak Tamil, 1507 speak Hindi, 572 speak English, 676 speak Tamil and Hindi, 286 speak Tamil and English, 270 speak Hindi and English, 114 speak Tamil, Hindi and English. Check whether the information is correct?

5. 3 Independence and Association of Data

Two attributes A and B are said to be independent if there is same proportion of A's amongst B's as amongst $\beta's$ or vice versa.

Two attributes A and B are independent if one of the following is true.

(1) $\dfrac{(AB)}{(B)}=\dfrac{(A\beta)}{(\beta)}$

(2) $\dfrac{(AB)}{(A)}=\dfrac{(\alpha B)}{(\alpha)}$

(3) $(AB)=\dfrac{(A)(B)}{N}$

(4) $(A\beta)=\dfrac{(A)(\beta)}{N}$

(5) $(\alpha\beta) = \frac{(\alpha)(\beta)}{N}$

(6) $(\alpha B) = \frac{(\alpha)(B)}{N}$

(7) $(AB)(\alpha\beta) = (A\beta)(\alpha B)$

Note :

(1) If $(AB) = \frac{(A)(B)}{N}$, we say that A and B are associated.

(2) If $(AB) > \frac{(A)(B)}{N}$, we say that A and B are positively associated.

(3) If $(AB) < \frac{(A)(B)}{N}$, we say that A and B are negatively associated.

To measure the intensity of two attribute A and B is called coefficient of association. Most commonly used coefficient of association is **Yule**'s coefficient of association or coefficient of colligation.

The Yule's coefficient of association is defined as

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

and the coefficient of colligation is defined as

$$Y = \frac{1 - \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}{1 + \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}$$

Note :

(1) If $Q = Y = 0$ then the attributes A and B are independent.

(2) If $Q = Y = 1$ then the attributes A and B are perfectly associated.

(3) If $Q = Y = -1$ then the attributes A and B are perfectly disassociated.

Example 5. 16

State and prove the relationship between coefficient of association and coefficient of colligation.

Solution :

Statement : The relationship between coefficient of association and coefficient of colligation is $Q = \frac{2Y}{1+Y^2}$.

Space for Hint

Space for Hint

Proof :

Let $x = \dfrac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}$

Then $Y = \dfrac{1 - \sqrt{\dfrac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}{1 + \sqrt{\dfrac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}$

$\Rightarrow \; Y = \dfrac{1-\sqrt{x}}{1+\sqrt{x}}$

$\therefore \; Y^2 = \left(\dfrac{1-\sqrt{x}}{1+\sqrt{x}}\right)^2$

$= \dfrac{1-2\sqrt{x}+x}{(1+\sqrt{x})^2}$

Thus $1+Y^2 = 1+\dfrac{1-2\sqrt{x}+x}{(1+\sqrt{x})^2}$

$= \dfrac{2(1+x)}{(1+\sqrt{x})^2}$

Hence $\dfrac{2Y}{1+Y^2} = \dfrac{2\dfrac{1-\sqrt{x}}{1+\sqrt{x}}}{\dfrac{2(1+x)}{(1+\sqrt{x})^2}}$

$= \dfrac{1-x}{1+x}$

$= \dfrac{1-\dfrac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}{1+\dfrac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}$

$= \dfrac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$

$= Q$

Therefore $Q = \dfrac{2Y}{1+Y^2}$.

Example 5. 17

Find the association between A and B when $N=1000$, $(A)=470$, $(B)=620$, $(AB)=320$.

Solution :

Given that $N=1000$, $(A)=470$, $(B)=620$, $(AB)=320$.

$$\text{Now } \frac{(A)(B)}{N} = \frac{470 \times 620}{1000}$$
$$= 291.4$$
$$< 470$$
$$= (AB)$$

(i.e.) $\frac{(A)(B)}{N} < (AB)$

Therefore attributes A and B are positively associated.

Example 5. 18

Find the association between A and B when $(AB)=90$, $(A\beta)=65$, $(\alpha B)=260$, $(\alpha\beta)=110$.

Solution :

Given that $(AB)=90$, $(A\beta)=65$, $(\alpha B)=260$, $(\alpha\beta)=110$.

Now

Attribute	B	β	Total
A	90	65	155
α	260	110	370
Total	350	175	525

$$\text{Now } \frac{(A)(B)}{N} = \frac{155 \times 350}{525}$$
$$= 103.33$$
$$> 90$$
$$= (AB)$$

Space for Hint

Space for Hint

(i.e.) $\frac{(A)(B)}{N} > (AB)$

Therefore attributes A and B are negatively associated.

Example 5. 19

In an examination in Tamil and English 245 of the candidates passed in Tamil, 147 passed in both, 285 failed in Tamil and 190 failed in Tamil but passed in English. How far is the knowledge in the two subjects associated?

Solution :

Let A be an attribute that a candidate passed in Tamil subject.

Let B be an attribute that a candidate passed in English subject.

Given that $(A) = 245$, $(AB) = 147$, $(\alpha\beta) = 285$ and $(\alpha B) = 190$.

The remaining class frequencies can be obtained from the following table.

Attribute	B	β	Total
A	147	98	245
α	190	285	475
Total	337	383	720

Now $\frac{(A)(B)}{N} = \frac{245 \times 337}{720}$

$= 114.67$

< 147

$= (AB)$

(i.e.) $\frac{(A)(B)}{N} < (AB)$

$\therefore$ A and B are positively associated.

(i.e.) the knowledge in the two subjects is positively associated.

Example 5. 20

Calculate Yule's coefficient of association between marriage and failure of studies.

	Passed	Failed	Total
Married	90	65	155
Unmarried	260	110	370
Total	350	175	525

Solution :

Let A be an attribute that denote marriage and

let B be an attribute that denote failure of studies.

Given that

Attribute	B	β	Total
A	90	65	155
α	260	110	370
Total	350	175	525

Therefore Yule's coefficient of association between marriage and failure of studies = Q

$$= \frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$$

$$= \frac{(65)(260)-(90)(110)}{(65)(260)+(90)(110)}$$

$$= \frac{7000}{26800}$$

$= 0.2612$

Space for Hint

Space for Hint

Example 5. 21

Investigate the association between darkness of eye colour in father and son from the following data.

Father with dark eyes and sons with dark eyes	78
Father with dark eyes and sons with not dark eyes	122
Father with not dark eyes and sons with dark eyes	96
Father with not dark eyes and sons with not dark eyes	704

Can we infer eye colour is hereditary?

Solution :

Let A be the attribute that denote dark eye to a father and let B be the attribute that denote dark eye to a son.

Thus we have the following

Attribute	B	β	Total
A	78	122	200
α	96	704	800
Total	174	826	1000

Therefore Yule's coefficient of association between marriage and failure of studies = Q

$$= \frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$$

$$= \frac{(78)(704)-(122)(96)}{(78)(704)+(122)(96)}$$

$$= \frac{43200}{66624}$$

$$= 0.648$$

Example 5. 22

From the figures given in the following table compare the association between literacy and unemployment in rural and urban areas and give reasons for the different if any.

	Urban	Rural
Total adult males	25 lakhs	200 lakhs
Literate males	10 lakhs	40 lakhs
Unemployed males	5 lakhs	12 lakhs
Literate and unemployed males	3 lakhs	4 lakhs

Solution :

Let A, B denotes literacy of males and unemployed males respectively.

Step 1 :

First we shall find the coefficient of association between A and B in urban.

Thus we have

Attribute	B	β	Total
A	7	3	10
α	13	2	15
Total	20	5	25

$$\text{Thus } Q = \frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$$

$$= \frac{(7)(2)-(13)(3)}{(7)(2)+(13)(3)}$$

$$= \frac{14-39}{14+39}$$

$$= -0.472$$

Space for Hint

Space for Hint

Step 2 :

First we shall find the coefficient of association between A and B in rural. Thus we have

Attribute	B	β	Total
A	36	4	40
α	152	8	160
Total	188	12	200

$$\text{Thus } Q = \frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$$

$$= \frac{(36)(8)-(152)(4)}{(36)(8)+(152)(4)}$$

$$= -\frac{320}{896}$$

$$= -0.357$$

Step 3 : from step 1 and step 2 it is clear that the association between the literacy and unemployed males in urban is greater than that in the urban.

Example 5. 23

160 plants are characterized as per the nature of the leaves and colour of the flower.

	Normal leaves	Abnormal leaves
White flowers	99	36
Red flowers	20	5

Examine the statement "the colour of the flowers and the nature of the leaves are independent".

Space for Hint

Solution :

Let A be an attribute that denote colour of the flower and

let B be an attribute that denote nature of leaves.

Given that

Attribute	B	β
A	99	36
α	260	110

Therefore Yule's coefficient of association between marriage and failure of studies = Q

$$= \frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$$

$$= \frac{(99)(5)-(20)(36)}{(99)(5)+(20)(36)}$$

$$= -\frac{225}{1215}$$

$$= -0.185$$

Thus the colour of the flower and nature of leaves are not independent but they are negatively associated.

Check Your Progress

(1) calculate the coefficient of association between intelligence father and son from the following data.

Intelligent fathers with intelligent sons	200
Intelligent fathers with dull sons	50
Dull fathers with intelligent sons	110
Dull fathers with dull sons	600

Comment on the result.

Space for Hint

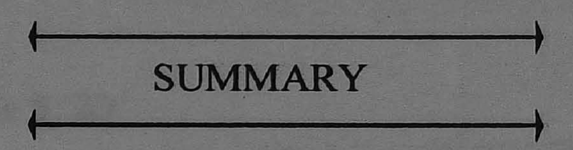
(2) From the following data compare the association between marks in Physics and Chemistry in two universities X and Y.

	X	Y
Total number of candidates	200	1600
Pass in Physics	80	320
Pass in Chemistry	40	90
Pass in Physics and Chemistry	20	30

(3) Of 500 students appeared for a competitive examination 350 were successful 280 had attended a coaching class and of these 220 came out successful. Does the coaching help in success?

(4) From the following information from the table, discuss the association between the colour of the skin and colour of the eyes..

Colour of skin	Colour of the eye	
	Black	Brown
Black	25	10
Red	12	38



SUMMARY

In this unit we learned that how to find the association between attributes, whether they are independent, positively associated or negatively associated. Also we have discussed the Yule's coefficient of association.